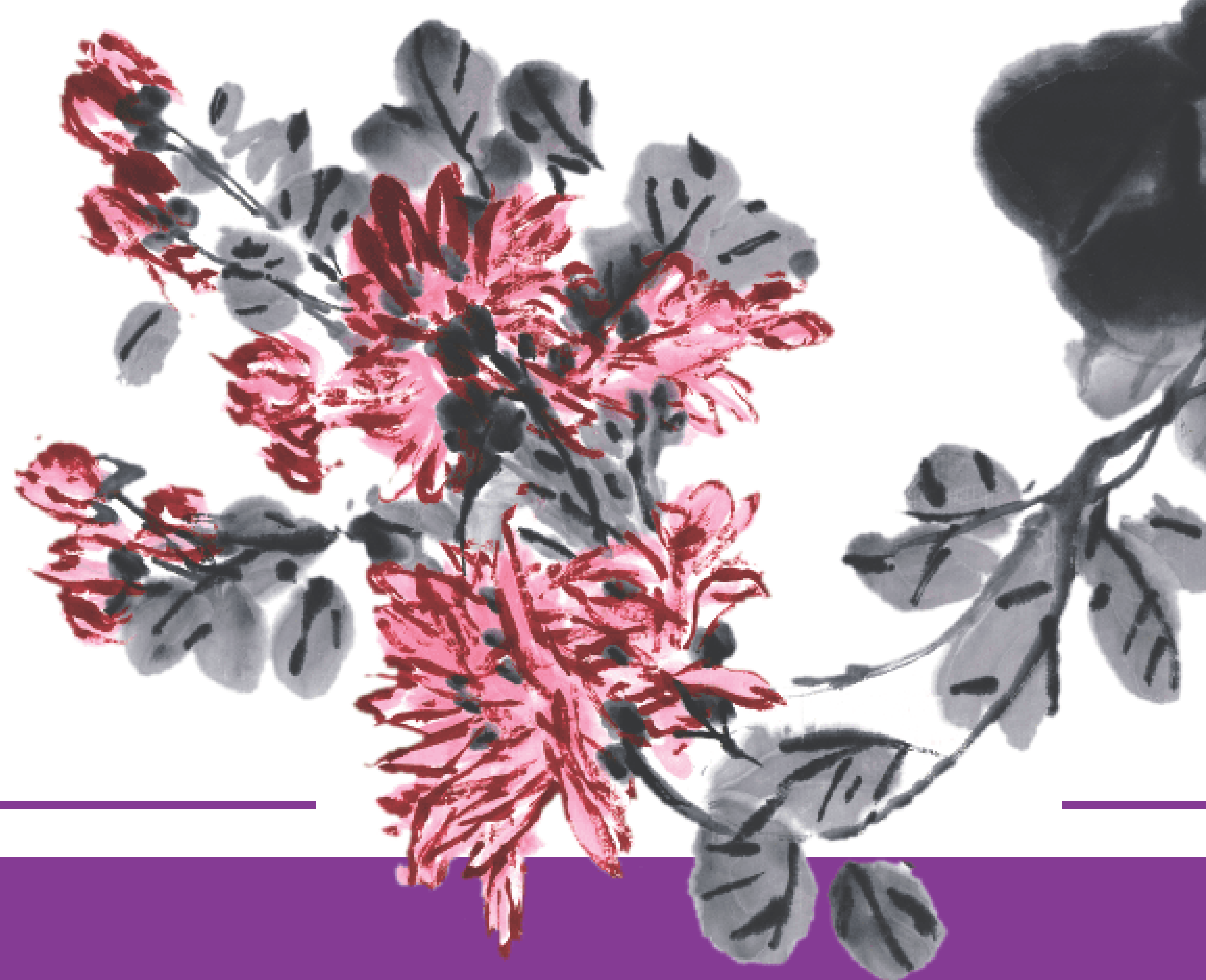


21世纪高等院校通识教育规划教材

DAXUESHENG XINXI JIANSUO SUYANG JIAOCHENG

大学生信息检索素养教程

王 冲　编著
Wang Chong



清华大学出版社

21 世纪高等院校通识教育规划教材

大学生信息检索素养教程

王　冲　编著

清华大学出版社
北　京

内 容 简 介

本书属于高等学校各个专业研究生和本科生的"信息检索素养课程"教学通用教材，内容包括三大部分：第一部分"信息检索素养基础知识篇"，第二部分"信息检索素养基本原理篇"和第三部分"信息检索素养实践应用篇"，共13章内容。本书较好地把现代信息检索素养知识的基础性与前沿性、原理性与实践性、全面性与主题性、引导性与启发性进行了贯通与融合。在基于大量信息检索专题、图表、实例及其数学理论依据进行充分阐述和说明的基础上，突出国内与国外、理论与实践紧密结合的信息检索素养教学要求。考虑到不同专业和不同层次学生的实际教学需要，教学内容组织依据循序渐进和主题性教学相结合的原则，可以适当选用部分章节组织教学。例如，针对计算机学科专业、图书情报学专业、信息管理专业本科生和各个专业的研究生层次学生，可以把第二部分"信息检索素养基本原理篇"作为重点来组织各个教学章节内容。

本书内容丰富、线索清晰、结构完整、语言精练、主题鲜明，是高等学校各个专业研究生和本科生的信息检索素养教学通用教材。既可以作为信息检索素养基础必修课教材，也可以作为部分专业和图书馆用户教育的选修课教材，同时可作为信息系统设计与开发、数据采集与挖掘、信息检索与咨询服务、图书情报机构等从业人员的学习与培训参考用书。

图书在版编目(CIP)数据

大学生信息检索素养教程 / 王冲编著. —北京：清华大学出版社，2017
(21世纪高等院校通识教育规划教材)
ISBN 978-7-302-46006-0

Ⅰ.①大…　Ⅱ.①王…　Ⅲ.①信息检索－高等学校－教材　Ⅳ.①G254.9

中国版本图书馆CIP数据核字(2016)第313719号

责任编辑：白立军　薛　阳
封面设计：傅瑞学
责任校对：梁　毅
责任印制：杨　艳

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦A座　　**邮　　编：**100084
社 总 机：010-62770175　　**邮　　购：**010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn,010-62795954
印 刷 者：北京富博印刷有限公司
装 订 者：北京市密云县京文制本装订厂
经　　销：全国新华书店
开　　本：185mm×230mm　　**印　　张：**34.75　　**字　　数：**674千字
版　　次：2017年1月第1版　　**印　　次：**2017年1月第1次印刷
印　　数：1～2000
定　　价：59.50元

产品编号：059891-01

前　言

在信息化社会越来越发达的今天，面对几何级数膨胀的海量信息资源，如何有效地检索、获取、评估、传播、共享和利用信息，成为了每个人重要的基本素养和能力要求，因为信息需求是每个人学习、工作、生活及其社会活动中十分重要而且迫切的需求。作为信息时代的大学生，需要重视信息检索素养的知识学习与能力培养。信息检索素养的理论知识学习与基本能力形成，不仅直接影响着大学生的在校学业表现，也较大程度上影响着他们今后的学习、工作与事业发展（例如终身学习、创新创业等持续性需要）。

大学生信息检索素养是大学生信息素养的核心内容之一，具有多学科交叉融合的特性。信息检索起源于图书馆学、情报学的信息检索原理与技术，早期直接服务于高校图书馆或社会公共图书馆的信息检索用户教育与技能培训，后来广泛应用于数据库研发与服务企业、搜索引擎等信息服务产业，在当今高速发展的计算机科学、软件工程、网络工程、通信工程、管理学、应用数学、统计学、语言学等多学科交叉融合的基础上，信息检索在数据挖掘、大数据处理等领域不断深化并发挥着日益强大的潜能。大学生信息检索素养教育正是基于这种时代背景和学科发展提出来的，也是面向大学生的传统信息素养教育和信息检索教育的不断深化与交叉融合的发展结果。

基于循序渐进和主题性教学原则，本书较好地把现代信息检索素养知识的原理性与实践性、全面性与主题性、引导性与启发性进行了贯通与融合。在基于大量信息检索原理与知识的专题、图表、实例、案例及其数学理论依据进行充分阐述和说明的基础上，突出国内与国外、基础与前瞻、知识与技能紧密结合的信息检索素养教学要求。考虑到不同专业和不同层次学生的实际教学需要，本教材属于高等学校各个专业研究生和本科生的“信息检索素养课程”通用教材，内容包括三大部分：第一部分“信息检索素养基础知识篇”，第二部分“信息检索素养基本原理篇”和第三部分“信息检索素养实践应用篇”。

本书逻辑清晰，内容丰富，结构完整。首先，从信息检索素养的基本概念、内涵、发展动因、特点、核心内容与能力表现、信息检索素养的评价标准以及信息化社会对大学生的信息检索素质需要出发，进一步论述信息检索与知识产权、信息检索与大学生学术不端行为、信息检索基础知识、信息检索方法与策略等内容来培养学生的信息检索意识、信息检

索道德与信息检索基础。第二，通过“信息检索的基础数学原理”的引入，使得信息检索有了更加严谨的逻辑论证，检索过程和信息需求的本质描述也更为精确，从而使得信息检索的理论与实践获得持续性的基础支撑。通过“文本分类与文本索引构建”、“图像信息检索”、“音频信息检索”、“视频信息检索”和“Web 信息搜索一般性原理”来构建大学生特别是研究生的信息检索基本原理知识。第三，通过“搜索引擎的检索应用”、“七大类特种文献信息资源检索”和“图书与学术期刊论文检索”的大量实例与检索案例来培养和锻炼大学生的信息检索素养实践技能。

本书教学内容的规划、组织与编著，是在作者讲授研究生“信息检索原理与应用”课程和本科生“大学生信息检索”课程的十多年教学改革与实践经验基础上逐步积累形成的。同时，在教材编著过程中，参考和借鉴了大量国内外专著、教材、学术期刊论文、学位论文、学术观点和典型网络数据库检索平台等成果，在此一并向他们表示真挚的谢意！

本书内容丰富、线索清晰、结构完整、语言精练、主题鲜明，是高等学校各个专业研究生和本科生的信息检索素养教学通用教材。既可以作为信息检索素养基础必修课教材，也可以作为部分专业和图书馆用户教育的选修课教材，同时可作为信息系统设计与开发、数据采集与挖掘、信息检索与咨询服务、图书情报机构等从业人员的学习与培训参考用书。

在本书编著过程中，得到桂林电子科技大学研究生院领导及教学督导委员会的关心与支持，获得“2016 年桂林电子科技大学研究生教育质量工程专项(YXYJ2900)”、“2016 年广西学位与研究生教育改革与发展专项(2016XWYJ12)”和“2015 年广西高等教育本科教学改革工程项目(2015JGA207)”的支持与资助。本书能够顺利出版，感谢清华大学出版社的大力支持与良好合作，感谢出版社编辑们的辛勤工作与付出！

本书主要基于循序渐进性教学与主题性教学相结合的编写原则，在大学生信息检索素养的原理性与实践性、全面性与主题性、引导性与启发性等方面难免有疏漏或不妥之处，恳请读者批评指正。

作　者

2016 年 7 月于桂林

目　录

第一部分　信息检索素养基础知识篇

第二部分 信息检索素养基本原理篇

第三部分　信息检索素养实践应用篇

第一部分

信息检索素养基础知识篇

信息检索素养可以描述为：善于根据问题分析自身的信息需求（例如学习或工作需要），进而确定信息来源并使用有效的检索或查找方法，及时地获取需要的信息；善于整理信息、分析评价信息，善于运用信息技术处理信息并用于解决问题；在信息的获取、处理、共享、使用的过程中具有良好的信息意识、信息道德和强烈的社会责任心，有一定创新、协作和服务精神。信息检索意识、信息检索技能和信息利用伦理道德是个体内在信息检索素养的外在表现，也是信息检索素养的基本要素。

第1章说明了信息检索素养的概念含义、发展动因、特点、主要内容与评价标准。同时说明了我国当代大学生信息检索素养的现状，阐述了进行信息检索素养教育与培养的必要性与重要作用。

第2章阐述了信息检索与知识产权，同时说明了知识产权的含义与内容。本章重点阐述了信息检索与利用的相关法律制度、信息检索与利用过程中的道德自律以及信息检索与利用同知识产权保护的相互影响。通过本章的学习，旨在培养大学生的信息检索道德和信息获取的相关法律知识。

第3章阐述了信息检索基本知识。包括检索的概念、信息检索的含义与类型、信息检索涉及的相关支撑领域、信息检索的前沿与热点问题。通过本章学习，旨在使读者总体把握信息检索的基本知识。

第4章旨在进行有关信息检索方法的知识学习,初步形成大学生必须的信息检索方法与技能性知识基础。内容包括:信息源及其加工层次类型、信息源及其物理载体类型、信息源的出版发行与共享类型、主要信息源类型的辨别、信息检索工具的基本功能与类型、信息检索途径、信息检索方法与策略以及信息检索质量与评价等内容。

需提示的是:在有限篇幅内,本篇不可能把“信息检索素养基础知识”进行全面概述与阐述(例如信息检索道德所涉及的知识产权与法律问题以及详细发展历史等内容),如因学习需要,可以查阅相关书籍;第4章所涉及的有关信息检索方法的原理性知识(例如信息检索的布尔逻辑组配系、构造高级检索表达式等),将在第二部分“信息检索素养基本原理篇”和第三部分“信息检索素养实践应用篇”中详细阐述。

第 1 章　大学生信息检索素养概述

当今世界，因为信息产业的经济总量超过了工业经济，也远远超过了农业经济，所以人类总体上已无可置疑地步入了信息时代，而且以惊人的速度、规模和爆发力不断改造和提升着现代工业、现代农业和现代服务业的快速进步。计算机技术、数据通信技术、多媒体技术等 IT 技术无时无刻不在深刻影响着我们每个人的学习、工作和生活。

根据中国互联网信息中心(http://www.cnnic.cn)于 2016 年 1 月发布的“中国互联网络发展状况(第 37 次)统计报告”显示：截至 2015 年 12 月，中国网民规模达 6.88 亿人，中国网民的人均周上网时长为 26.2 小时，互联网普及率达到 50.3%，半数中国人已接入互联网，其中有 90.1%的网民通过手机上网。网民数量的激增和旺盛的市场需求推动了互联网领域更广泛的应用发展热潮。1.10 亿网民通过互联网实现在线教育，1.52 亿网民使用网络医疗，9664 万人使用网络预约出租车，网络预约专车人数已达 2165 万人，网上支付用户规模达 4.16 亿人，全国开展在线销售的企业比例为 32.6%，开展在线采购的企业比例为 31.5%，我国网站总数为 423 万个，中国网页数量为 2123 亿个。

在信息时代的今天，面对几何级数增长的海量信息资源，如何有效地检索、获取、评估、传播、共享和利用信息，成为了每个人重要的基本素养和能力要求。作为信息时代的大学生，需要重视信息检索素养的知识学习与能力培养。信息检索素养的理论知识学习与基本能力形成，不仅直接影响着大学生的在校学业表现，也较大程度上影响着他们今后的学习与发展(例如终身学习、创新创业等持续性需要)。

信息检索素养是一个得到持续和广泛研究的课题，在社会信息化不断提升的今天，对大学生而言其重要性更为凸显。据我国图书情报学领域专家赖茂生的研究：本科生检索方法手段单一，使用搜索引擎查找生活、娱乐类信息，对搜索引擎易用性的判断高于 OPAC 和数据库，无论是检索字段的使用还是对检索结果的判定，其所凭借和依据的字段或内容均很少；与受过专业训练的信息管理专业的大学生(含研究生)相比，其他专业的大学生在对检索结果的甄别能力上存在显著差异，大学生对特定的信息检索系统(如搜索引擎)有着较强的偏好，但是对信息检索系统所提供的辅助手段(如高级检索语法)的使用却不尽如人意。

1.1 信息检索素养概述

1.1.1 信息检索素养的基本概念

信息检索素养(information retrieval literacy)的内涵与外延与“信息素养”的概念含义较为相近。信息素养(information literacy)一词,最早是由美国信息产业协会(IA)主席保罗·泽考斯基在1947年提出的:“利用大量的信息工具及主要信息源使问题得到解答的技术和技能。”他强调信息素养是一种信息检索的信息查询、获取与利用以解决问题的技能和能力,体现了对于信息社会每个公民的一项基本能力要求。由于“信息素养”概念从产生之初就与“信息检索素养”概念含义相近,所以在很多正式场合(专著、论文、会议文献或课堂教学的教案资料与教材等)“信息检索素养”与“信息素养”是等同对待的,尽管后来的发展对“信息素养”概念有一定拓展和延伸。

信息检索素养的概念含义大多基于图书馆学、情报学的学科角度,而信息素养的含义大多基于社会学的角度。美国图书馆协会给予的定义:“一个有信息素养的人,必须能够确定何时需要信息,并且具有检索、评价和有效使用所需信息的能力。”它简要地概括了信息检索素养的主要内容与完整过程。

大学生信息检索素养的含义,一方面它体现着一种终身学习的理念和自主学习的能力,这也是我国大力提倡的教育理念和目标;另一方面它表现为搜集信息、解决问题的能力,这不仅仅表现为检索信息能力,更是对创造性思维的考验,同时它要求具备道德法律意识,在法律允许、道德约束下进行信息检索与利用活动。此外,在当今社会,信息检索素养不仅仅是一个人解决问题的能力,更重要的是一种潜在的思想、意识和个人素质。

“信息检索素养”指有能力从各种不同信息源(Web数据库、图书馆资源库、专门检索工具或引擎平台等)中查询、获取、评价和使用信息。信息检索素养可以概括为一个人在查找与获取信息、处理和共享信息并利用信息方面的知识和能力品质。信息检索素养既是个体查找、检索、分析信息的信息认识能力,也是个体整合、利用、处理、创造信息的信息应用能力。具体描述为:善于根据问题分析自身的信息需求(例如学习或工作需要),进而确定信息来源并使用有效的检索或查找方法,及时地获取需要的信息;善于整理信息、分析评价信息,善于运用信息技术处理信息并用于解决问题;在信息的获取、处理、共享、使用的过程中有良好的信息意识、信息道德和强烈的社会责任心,有一定创新、协作和服务精神。

因此,信息检索素养包含了检索技术和人文精神两个层面的意义:在检索技术层面

上，信息检索素养反映的是人们利用信息检索的意识和能力；在人文层面上，信息检索素养反映了人们利用信息时表现出来的品质和修养（例如信息产权意识、信息安全意识、不良信息过滤与免疫、网络暴力抵制与防护、杜绝抄袭与剽窃学术不端行为等）。大学生要想在信息社会中更好地生存和发展，不断提高自身的学习、工作和生活效率就必须具备良好的信息检索素养。获取、评价、共享与利用信息资源的知识和能力，已经在大学生的学习与研究、生活与娱乐、实践与工作等环节发挥着越来越重要的作用。

可以从广义和狭义两个角度来进一步理解信息检索素养的概念含义。

广义而言：信息检索素养是个人内在综合修养的一个重要方面。它外在表现为个体在为实现认知而进行的信息活动中所表现出来的文化素养、信息检索意识、信息检索技能和信息利用伦理道德观念的总和。简而言之，文化素养、信息检索意识、信息检索技能和信息利用伦理道德是个体内在信息检索素养的外在表现，也是信息检索素养的四大基本要素。这里的文化素养有两方面含义：一方面是检索知识的学习；另一方面是指个体在工作、生活中对所面临信息需求的问题或任务及相关信息的认识和处理。

狭义：如果依据学习的信息加工理论，把认知看做是信息的加工，它是转换、简约、储存、提取和使用等活动输入的过程，那么信息检索素养就是在获取、运用、加工信息，生成、创造、表达新信息的过程中所表现出来的综合能力。

总之，信息检索素养是一个含义广泛的综合性、发展性的概念，信息检索素养不仅包括利用信息工具和信息资源的能力，还包括获取识别信息、加工处理信息、传递与创造信息的能力，更重要的是以独立自主学习的态度和方法、以批判精神以及强烈的社会责任感和参与意识，并将它们用于实际问题的解决和进行创新性思维的综合的信息能力。

1.1.2　大学生信息检索素养的内涵

信息检索素养是很多领域的研究重点，这与信息素养教育能够直接促进个人乃至社会的发展有关。图书情报领域是信息检索素养研究的一支主要力量。例如，美国图书馆协会提出《高等教育信息素养能力标准》（*Information Literacy Competency Standards for Higher Education*），该标准从五个方面来揭示信息检索素养的内涵。

（1）确定信息需求的本质和范围。

（2）优质高效地获取所需信息。

（3）客观地评价信息和信息源，并将所选取信息纳入其知识库和价值系统。

（4）使用信息完成给定的任务。

（5）理解与信息使用和获取相关的经济、法律和社会议题，并合理合法地使用信息。

信息检索素养的另一内涵框架是由美国 Eisenberg 和 Berkowitz 提出的 Big6 能力，包括任务定义、信息查寻策略、定位与获取信息、信息使用、综合和评估六个方面。我国图书情报界的信息检索素养研究侧重于问题解决和信息服务提供；教育技术领域则关注信息科学知识、信息检索能力、信息检索情感意识和信息伦理道德四个方面。大学生信息检索素养的内涵可以包括以下几个方面。

(1) 信息检索意识。信息检索素养教育最重要的一点是培养大学生的信息检索意识，即要求大学生具有一种使用计算机与其他信息技术来解决自己学习、工作和生活中信息需求问题的意识。

(2) 信息检索伦理修养。大学生能够遵循信息应用的伦理道德规范，不从事非法活动，同时也知道如何防止计算机病毒和其他计算机犯罪活动，在法律法规允许的范围内合理合法地检索与利用信息资源。

(3) 信息检索技术知识。掌握信息检索技术的原理、名词术语与基本应用，了解信息检索技术发展与作用，具有一定的信息检索技术知识，把握信息检索技术的发展与应用。

(4) 具有一定的信息检索能力。即查询、评价和利用信息以提高学习、工作和生活效率的能力。能利用信息技术获取自己所需要的信息，评价和分析所得到的信息，并有效地利用在自身的学习、工作和生活之中。

1.1.3 信息检索素养的发展动因

(1) 一种个体的基本能力素养。当今信息量的几何级数膨胀和海量信息中信息质量的不确定性造就了信息超载的局面，对人们认识、检索、使用和评价信息的能力形成了挑战。人们在社会生活的各个方面面临着不同种类的、数量巨大的信息把握与选择，为保证其真实性、完整性和安全性，必须以有效的手段去获取、利用和鉴别信息，这种能力来自信息检索素养教育。美国教育技术 CEO 论坛 2001 年第 4 季度报告提出 21 世纪人才的重要能力标准包括的五个方面：基本学习技能（指读、写、算）、信息素养、创新思维能力、人际交往与合作精神、实践能力。信息素养是其中的一个重要要素之一。大学生的信息检索素养要求，比较典型的有来自美国高校和研究图书馆协会 CRAL 特别工作组，他们提出高等院校学生应具备的信息检索素养有六大指标：①确定所需信息的范围；②有效地获取所需的信息；③鉴别信息及其来源；④将检出的信息融入自己的知识基础；⑤有效地利用信息去完成一个具体的任务；⑥了解利用信息所涉及的经济、法律和社会问题，合理、合法地获取和利用信息。六大指标下包括 22 个二级指标和 86 个可测评的科目。

(2) 一种个体的综合信息能力要求。信息检索素养包括广泛的概念（例如信息检索

的数学原理、计算机数据挖掘、云计算与大数据处理、互联网法规等），和许多学科相关，包含人文的、技术的、经济的、法律等诸多知识背景。IT 技术支撑信息素养，是信息检索素养的一种有力技术工具。信息检索素养这种信息能力，包括信息智慧、信息道德、信息意识、信息觉悟、信息观念、信息潜能、信息心理等多个方面，它是一种了解、搜集、评价和利用信息的知识结构，需要借助信息技术、依靠完善的查询与利用方法、通过鉴别和推理来完成实际的信息应用与再创新。

1.1.4　信息检索素养的特点

同信息检索素养的含义密切相关的是信息检索素养特点，明确其特点既有助于进一步把握信息检索素养的含义本质，也有助于形成信息检索素养培养的明确对策。

1. 信息检索素养的普遍性

信息检索素养的普遍性是指在信息社会中，信息检索素养普遍存在于社会的各个领域，属于每一个人的一种广泛的基本素养。信息检索素养普遍性之所以存在的根本原因是信息检索与获取需求无时不有、无处不在。

首先，在信息社会中，信息资源日益成为社会各领域中最活跃、最具有决定意义的因素，是一种普遍存在的重要的支撑性资源，基于知识和信息的新经济形态已经形成庞大的规模，信息产业成为国民经济的支柱产业，信息技术的飞速发展使“21 世纪是知识与信息的时代”成为共识，信息的财富意识业已形成，善于拥有信息资源就能够利用信息开发、设计出所需要的产品并占领市场，能够获取巨大的经济利益，最典型的就是目前流行的“互联网＋”对各个行业生态的大力渗透。其次，信息技术的发展使知识的载体发生了根本性的变化。承载知识的是“比特”，即以二进制形式存储的数字媒体，其基本特性在于无限的再生性和不受任何限制的传播性，人们可以迅速地获取大量所需的知识和信息，出现所谓“知识大爆炸”的现象。据联合国教科文组织的统计：人类近 30 年来所积累的科学知识占有史以来积累的知识总量的 90％，而在此之前的几千年中所积累的科学知识只占 10％。再次，信息技术的应用深入到社会生活的各个领域，成为人类生活的一部分，信息检索素养已成为信息社会文明人应该具有的一种基本素养，是与读、写、算一样同等重要的、终身有用的基础能力。它没有年龄、职务、地域、时间上的区分，没有绝对权威，人们可以通过学校教育或自学来不断培育和提高信息检索素养。

2. 信息检索素养的层次性

信息检索素养的层次性是指在信息社会中，由于人们与信息技术应用的密切程度与实际信息需求的层次不同而具有不同要求的特点。信息检索素养的普遍存在，使具有良

好的信息检索素养成为信息社会对所有人的基本要求。但信息技术本身是一种高度知识化的技术，因此依据使用者与信息技术关系的密切程度不同和实际工作、学习、生活对信息需求层次的不同，信息检索素养可分为不同的层次与要求。

首先是公民基本型信息检索素养。在信息社会中，任何人都不可避免地与信息技术的应用联系着，这就要求所有的公民都应具有最基本的信息检索素养，这是对所有公民的要求，也是学校教育阶段所应培育的学生综合素质的一个重要组成部分。要通过学校教育培养他们对信息技术的兴趣和意识，掌握信息技术的基本知识和技能，了解信息技术的发展与应用对人类社会的深刻影响，培养学生良好的信息能力，教育学生负责任地使用信息技术，培养学生把信息技术作为支持终身学习和合作学习的手段，为适应信息社会的学习、工作和生活打下必要的基础，使他们成为信息社会的“合格网民”。

其次是职业操作型信息检索素养。作为信息技术应用人员所需要的信息检索素养是在公民信息检索素养的基础上建立起来的。他们通常要较为系统地了解信息技术的工作原理；具备通用工具软件的应用能力，并能按照职业与分工的要求，对某一类工具软件比较熟悉，掌握该软件所具有的各种特殊信息与数据的意义；具有较强的信息应用能力，能够充分发挥软件工具的功能，制作与开发出与本职业相关的各种各样的信息检索数据库产品。

再次是专业研究型信息检索素养。作为信息检索系统（例如专门学习型数据库、搜索引擎等）的开发设计人员，他们把信息检索系统的开发作为自己的职业或个人爱好。通常要求他们具有十分强烈的信息意识，具有较高的信息产权观和信息安全观；具有高度的信息伦理道德修养；熟悉信息检索与服务系统的工作原理与技术实质。作为信息检索与服务系统的开发设计人员，尤其是在信息能力方面要求更高，能够熟练应用各种通用工具软件与编程语言，掌握检索系统所具有的各种信息检索算法与特殊意义。同时更加强调在利用信息技术系统中的信息理解、信息选择、信息批判、信息收集、信息处理、信息生成、信息表达等方面的能力，并具备较强的程序设计与系统设计能力，从而能够不断开发出新的信息检索与服务产品，推动信息检索的技术性发展。

3. 信息检索素养的实践性与操作性

信息检索素养的实践性与操作性是指信息检索素养的学习与培育、提高与评价过程的最终体现都在于人们对于信息检索与利用的实践与操作上。首先，信息检索素养的学习与培育必须通过大量的实际操作来锻炼。就信息知识的掌握而言，只有通过具体的操作，把抽象的知识具体化，把深奥的信息检索技术理论化为实践行动，才能使人们对知识和有用信息有更深刻的认识与理解；较强的信息意识和信息能力只有通过不断操作与实

践,才会逐步提高,从而形成捕捉信息的敏锐性、筛选信息的果断性、检索和评估信息的准确性、交流信息的自如性和应用信息的独创性;正确的信息伦理道德也只有在信息检索活动的不断应用过程中才能发现问题与提出解决问题的办法,从而形成知-情-信-意-行的良好的信息伦理道德修养。其次,信息检索素养的评价集中表现在具体的使用与操作上。信息意识的强与弱,要看个体在实际操作中敢不敢、想不想使用信息技术,对信息检索技术的使用是否热心与积极;信息伦理道德的好坏,要看在实践中能否遵守各项法律法规,是否遵守网络文明公约,是否尊重他人的劳动成果等;判断一个人信息知识与能力的高低依据的是他能够知道多少,运用信息检索技术解决实际问题的水平如何。

4. 信息检索素养的发展性

信息检索素养的发展性是指随着信息技术的不断发展,人们的信息检索素养必然会不断地提高与发展,永无止境。从信息技术的发展来看,计算机信息技术的发展与普及尽管只有几十年的时间,但它已经历了电子管计算机、晶体管计算机、集成电路计算机以及大规模集成电路计算机的发展阶段。尤其是随着信息社会的来临,信息技术的发展日新月异,人们对信息检索素养的要求也越来越高,内涵越来越丰富,由最初强调计算机检索原理的程序编程能力,到包括信息意识、信息能力与信息知识、信息伦理道德等各个方面,并且随着信息技术的不断发展,其内涵将会不断发展。人们已经掌握的知识和技术很快就会被新的功能更强的技术与软件所取代。为适应社会的发展,人们将不断学习新的知识与技术,这不仅将有力地促进人们信息检索素养的发展和提高,而且通过信息技术教育还将有效地促进个体信息素养的全面提高。

1.2　信息检索素养的主要内容

信息检索素养主要包括信息检索意识、信息检索能力、信息检索道德、信息知识、信息观念、信息心理等方面,而信息检索意识(想到没想到,信息需求的有效获取意识)、信息检索能力(相应技能的会不会)、信息检索道德(信息共享与利用的合规合法性)又是其中的主要方面。

1.2.1　信息检索意识

信息检索意识是指人的头脑对信息需求满足及其检索原理的既抽象又概括的认识。信息检索意识是信息检索素养的前提与基础。信息检索意识主要表现在对需求信息具有高度的敏感性和积极主动的主动性。一个人的信息检索意识强,就能通过蛛丝马迹,自觉

地捕捉到任何有价值的信息资源来满足信息需求。信息检索意识的教育,主要就是培养大学生具有正确的信息需求观念、强烈的专业信息需求和持久的信息注意力。高校要注重培养大学生善于观察的习惯,除了关注自己的专业学科以及交叉学科信息外,还要及时发现和掌握最新动态信息,以便快速、准确而全面地获取和利用所需信息,这是创新人才的必备意识之一。在信息已成为经济和社会发展最为重要的战略资源的形式下,正视并重视信息的价值,做到充分有效地利用信息资源。

1.2.2 信息检索能力

信息检索能力是信息检索素养的核心内容,它是指个体能否依照自身的信息需求去捕捉与发现、查询与评价、选择与整合、吸收与利用信息,能否对信息进行加工并在获得信息的基础上进行创新性应用。

信息检索能力具体包括以下几方面。

(1) 自主、有效地运用各种检索工具和信息资源去查找和收集所需信息。

(2) 对收集和检索到的信息进行评价。

(3) 对已获取的信息进行整理、选择和整合。

(4) 将获得的信息纳入自己的已有知识体系中,即吸收信息。

(5) 对信息检索知识进行不断学习,对已获得的信息进行深入分析并用于实际问题解决,以最终达到满足信息需求的目的,同时创造并生成可能的新的相关信息。

信息检索能力表现过程如图 1-1 所示。

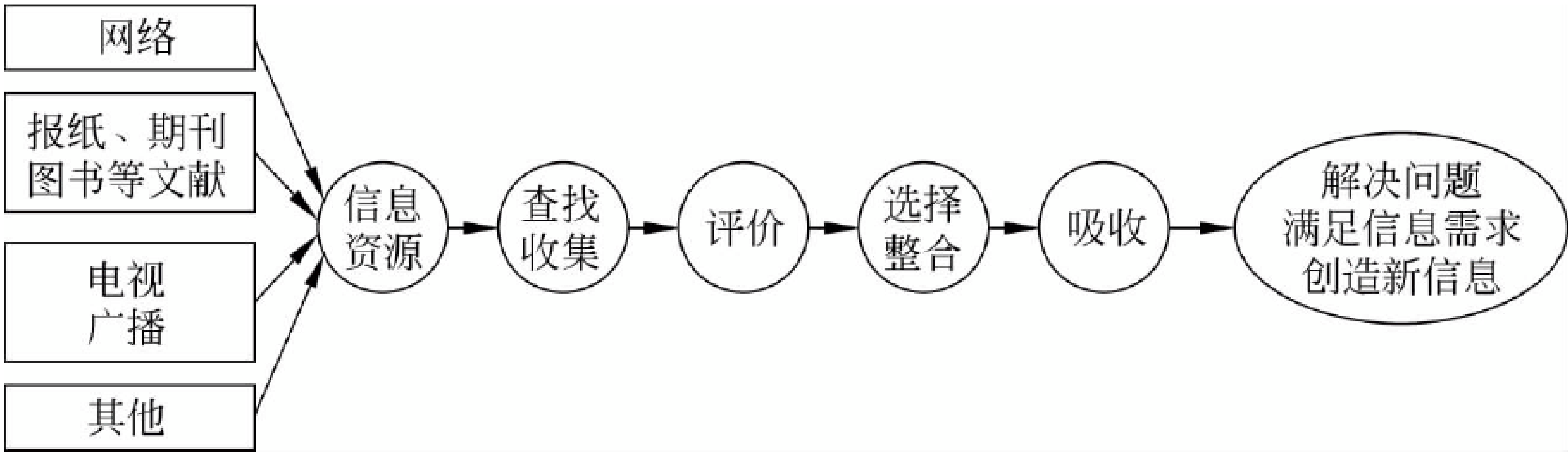


图 1-1 信息检索能力表现过程

1.2.3 信息检索道德

信息检索道德是信息检索素养的灵魂,信息检索道德在静态上是指个体在信息检索与利用活动过程中所应当遵循的道德行为规范,动态上则是表现为个体在进行信息活动

时自觉遵守法律和道德规范。信息检索道德调节着信息创造者、信息服务者、信息使用者之间的关系，规范着人们自身的信息检索行为，它是个体在信息活动中自觉承担社会责任的表现，包括不制造、传播和消费不良信息，不侵犯他人的知识产权、商业秘密、个人隐私，自觉坚持公正、平等、真实的原则，自觉抵制不良、恶意信息并积极与违法信息活动做斗争。尽管信息检索道德缺少实际可操作性的评价标准，但鉴于信息检索素养在某种意义上也是一种人文素养，它决定着个体的信息检索行为与利用互动是否能对自己、他人和社会产生积极作用。因此，树立良好的信息检索道德是有效预防和治理信息环境污染、避免抄袭、窃取、信息欺诈、网络暴力和信息破坏等信息检索与利用过程中道德失范的根本。

1.3　信息检索素养的评价标准

1.3.1　有信息检索素养的人

1989 年，美国图书馆协会和美国教育传播与技术协会提交了一份《关于信息素养的总结报告》，提出有信息素养的人必须：①认识到何时需要信息；②能够评价和使用所需的信息；③有效地利用所需的信息。有信息素养的人最终是指那些懂得如何学习的人，懂得如何学习是因为他们知道如何找到信息，知道如何利用信息。该报告开创了研究与评价“有信息检索素养的人”的先河。

1900 年，美国国家信息素养论坛在年度报告中提出有信息素养的人应是：①了解自己的信息需求；②明确所需信息的正确和完整是制定明智决策的基础；③能在信息需求的基础上系统阐述问题；④具有识别潜在信息源的能力，能制定成功的检索策略；⑤能检索信息资源，包括利用以信息为基础的信息技术或其他技术；⑥具有评价信息的能力，能为实际应用而对信息进行组织；⑦具有将新信息结合到现存的知识体系中的能力；⑧能采用批判性思维，利用信息并解决问题。该报告对研究与评价“有信息检索素养的人”进行了拓展和进一步发展。

Doyle 在《信息素养全美论坛的终结报告》中定义一个具有信息检索素养的人，他应该具有：①认识到精确的和完整的信息是做出合理决策的基础；②确定对信息的需求，形成基于信息需求的问题；③确定潜在的信息源，制定成功的检索方案；④从包括基于计算机和其他的信息源获取信息，评价信息，组织信息用于实际的应用；⑤将新信息与原有的知识体系进行融合以及在批判性思考和问题解决的过程中使用信息。

全美图书馆协会和美国教育传播与技术在《信息能力：创建学习的伙伴》一书

中，从信息素养、独立学习和社会责任三个方面，提出了学生的九条信息检索素养综合性标准。

1. 信息素养

（1）有信息检索素养的学生能有效地和高效地获取信息。

（2）有信息检索素养的学生能批判性地评价信息。

（3）有信息检索素养的学生能准确地、创造性地使用信息。

2. 独立学习

（1）独立的学习者要有信息检索素养，并能探求与个人兴趣相关的信息。

（2）独立的学习者要有信息检索素养，并能评价文献和其他信息的创造性表达。

（3）独立的学习者要有信息检索素养，并能力争在信息查询和知识的产生中做到最好。

3. 社会责任

（1）对学习团体和社会做出积极贡献的学生具有信息检索素养，并能认识信息对民主社会的重要性。

（2）对学习团体和社会做出积极贡献的学生具有信息检索素养，并能实践与信息和信息技术相关的合乎道德的行为。

（3）对学习团体和社会做出积极贡献的学生具有信息检索素养，并能积极参与小组的活动来探索和产生信息。

1.3.2 信息检索素养评价标准的必要性

1. 有助于评价个人的信息检索素养能力

信息检索素养能力作为信息社会人们的一项基本技能，涵盖很多方面的内容，包括信息检索意识、信息检索能力和信息检索伦理道德等。要对内涵丰富的信息检索素养能力进行评价，简单以课程测试、问卷调查等单一的评价方式进行，既不科学也不符合学科特征。因此，要科学地评价个人的信息检索素养，应采取多项目、多途径、多形式、动态与静态相结合的方法，灵活有效地进行评价。但要采取多种方式进行评价必须依赖一定的评价标准，只有根据统一的评价标准，才能保证实际的评价活动得以有效开展。

2. 有助于信息检索素养教育的规范化进行

信息检索素养评价标准之于信息检索素养教育的重要性日益得到图书馆学与情报学界、教育学界人士的认同。尤其在国外，从 20 世纪 70 年代已经开始研究，目前已经形成许多较为完善的信息检索素养标准。这些标准为进一步开展信息检索素养教育打下了坚

实的基础,使得信息检索素养教育在全世界范围规范化地开展,各国根据标准实施了一系列引人关注的信息检索素养教育项目。我国因信息素养评价标准的缺失,信息检索素养教育至今基本上还是各行其是,没有明显的特色和规范。只有进行信息检索素养评价标准研究,制定相关的信息检索素养评价标准,才能有针对性地制定相关课程计划和培养方案体系,从而改进信息检索素养教育现状,使得信息检索素养教育能够在较为规范化下的环境中进行。

1.3.3　大学生信息检索素养评价标准

信息检索素养是大学生终身学习的基础,适用于所有学科、所有学习环境和所有教育形式,因此高等教育应以培养大学生信息检索素养为重要内容之一。

美国大学与研究图书馆协会(American College and Research Libraries,ACRL)的《高等教育信息素质能力标准》提供了个人信息检索素质的能力架构。这个标准涵盖了大学各个年级的要求,可以借鉴参考。如表 1-1 所示。

表 1-1　高等教育信息检索素养评价标准

一级评价标准	二级评价标准
一、能确认信息需求本质与范围	1. 界定信息需求。 2. 知道辨识不同类型与媒体形式的信息资源。 3. 考虑取得所需信息的成本和效益。 4. 重新评估所需信息的特性与范围
二、能有效地获取所需的信息	1. 选择适当的检索方法和信息检索系统,以取得所需信息。 2. 构建有效的检索策略。 3. 利用上网或亲访等各种不同的方法,取得所需信息。 4. 必要时,重新界定检索策略。 5. 摘要、记录、管理信息资源
三、能批判地评估信息资源,并将其纳入自己的知识库与价值体系	1. 从所搜集的信息整合中,概要陈述主要概念。 2. 建立适当的准则,以评估信息与资源。 3. 综合重要概念,以构建信息观念。 4. 将新旧知识加以比较,以获得其价值、矛盾或独特之处。 5. 判断新知识对个人价值系统的影响,并调和其间差异。 6. 经由与他人和专家学者的互动,以验证诠释所得信息。 7. 判断是否要修正最初的查询疑问

续表

一级评价标准	二级评价标准
四、能有效地使用信息以达到个人或团体的特定目标	1. 利用信息和原有信息，以提高绩效。 2. 修正创作过程。 3. 有效地与他人分享创作成果
五、能了解信息使用的经济、法律与社会问题，并合理合法使用信息	1. 了解与信息、信息科技相关的伦理、法律与社会经济课题。 2. 遵循信息获得和使用的相关法律、法规、政策和各种约束。 3. 呈现创作成果并适时向信息来源致谢

1.4 我国当代大学生的信息检索素养现状

1.4.1 信息检索意识较弱

几乎所有的大学生都能意识到在信息化的今天有很多有用的信息，但是大多数学生无法主动发现有效的专业信息与社会信息同自身的学习、工作和生活紧密联系起来，通过信息资源的掌握和有效的信息检索活动来找到解决实际问题的有效途径。例如大量的网络课程学习网站、网络学习平台、学习与研究型数据库、各类慕课学习平台、各类虚拟实验平台等，对于大学生的课程学习有很好的帮助作用，但是学生的注册量与访问量并不高。

1.4.2 获取信息的检索能力不强

大多数学生获取信息的时候通常借助于一般的网络搜索引擎，没有掌握相应的信息检索的方法与技巧。例如，大多数网络数据库和搜索引擎都有高级检索功能（或专业检索功能），大多数学生不用或根本不知道怎么用。常常在不明确和细化信息需求的情况下，频繁使用一般检索功能或初级检索功能，这样表现出来的信息检索能力几乎与小学生的信息检索能力近似。

1.4.3 加工与利用信息的能力较差

大多数学生利用信息工具对信息进行加工处理的能力还处于较低的水平，只是停留在文字处理、上网浏览信息、简单的信息搜索、下载或截图、收发邮件等这些初级应用上，对信息的分析、筛选以及利用信息解决实际问题的能力还有待进一步加强。

1.4.4 信息道德和信息法规意识急需培养

只有少数的学生能够了解与信息获取和利用相关的法律法规、道德规范，对网络环境

下的“知识产权”、“个人隐私”、“信息安全”等方面的信息保护与信息防范意识薄弱，大学生对信息道德和信息法规内容的认识和了解不够全面，从而使恶意病毒、信息泄露、信息诈骗、网络暴力、低俗游戏、色情视频、不当言论等不安全或网络违法行为频频发生在大学生身上。举一个例子，一个突出的问题是很多大学生利用微博、微信、即时通信软件等社交媒体分享和暴露了自己、朋友甚至家人的很多隐私信息，自己还“乐在其中”，而不知“福兮祸之所依”。

1.5　大学生信息检索素养教育与培养的意义

21 世纪是信息化时代，信息爆炸对人的综合素质提出了日益严峻的挑战，国民是否具有良好的信息检索素养成为影响综合国力的一个重要方面。“百年大计，教育为本”，高校作为人才培养的重要基地，对普及信息检索素养教育，提高人才素质具有重要意义。为适应信息时代的发展变化，大学生应当具备信息检索素养的基本要求，信息检索素养也成为了国民综合素质的一个基本评价标准。信息检索素养是信息社会高等教育的重要内容，信息检索素养教育是培养学生了解信息资源、处理信息、有效利用信息、遵守信息道德规范的活动。加强信息检索素养，对于促进学生的学习效率、科研能力、创新创业实践与数字化生活质量等，具有十分重要的意义。

1.5.1　信息化社会对大学生的信息检索素质需求

人类进入 21 世纪，也进入了信息时代和知识经济时代。计算机技术、通信技术和网络技术的飞速发展，特别是因特网在全球的迅猛发展，标志着人类已经进入了一个全新的发展阶段即信息化社会。信息资源已成为信息化社会赖以生存和发展的重要资源，成为促使社会、经济和科学技术发生变革的主导因素。信息正以几何级速度骤增，引发了“信息爆炸”，使人们在进行学习、工作、生活和科学研究时，都面临着正确信息选择与合理利用的现实问题。通过网络、媒体、图书馆、社会、学校等提供、传播和交流的各种信息，形式多样且内容复杂，有文本、图像、视频、音频、动画等，大多数都是未经过过滤和筛选的，这就给人们选择、评价、理解和利用信息带来了新的挑战。信息在本质上具有不确定性，在量上具有无限扩展性，激增的信息量并没有让人们同步增加有效利用信息的能力，这就成为了面对信息时代人们生存立足的现实挑战。信息时代的高等教育完全不同于传统的高等教育，两者最根本的区别是从以教师为中心，以全面教育为主的教学模式，转变为以学生为中心、以个性化学习为主的模式，学生也是信息的生产者、传播者与合作者。

大学生必须适应信息化的社会环境，熟悉并掌握各种现代化信息资源的方式，具备发掘、获取所需信息的能力，具备高素质并学会运用信息技术手段检索与利用信息。信息检索素养较低或者缺失的学生，就不是一名合格的大学生。开展大学生信息检索素养教育是信息时代发展的需要，也是信息社会中人们的基本通行证。

我国教育信息化开展得如火如荼，对学生的信息检索素养教育也提出了更高的要求。我国教育信息化十年发展规划(2011—2020年)明确提出：着重解决国家教育信息化全局性、基础性、领域共性重大问题，实施“中国数字教育2020”行动计划，在优质资源共享、学校信息化、教育管理信息化、可持续发展能力与信息化基础能力五个方面，取得实质性重要进展。实施优质数字教育资源建设与共享是推进教育信息化的基础工程和关键环节。到2015年，基本建成以网络资源为核心的教育资源与公共服务体系，为学习者可享有优质数字教育资源提供方便快捷服务，建设各级各类优质数字教育资源。针对学前教育、义务教育、高中教育、职业教育、高等教育、继续教育、民族教育和特殊教育的不同需求，建设20 000门优质网络课程及其资源，遴选和开发500个学科工具、应用平台和1500套虚拟仿真实训实验系统。整合师生需要的生成性资源，建成与各学科门类相配套、动态更新的数字教育资源体系。公平、均衡、质量、创新、灵活、个性的教育信息化目标正在我国大力推进。具有良好信息检索素养的学生将真切感受到“无处不在的学习，无处不在的教育”。

2015年我国政府工作报告中，李克强总理首次提出“互联网＋”行动计划，以此推动移动互联网、云计算、大数据、物联网等与现代工业、农业、金融、教育、健康、商务等各行各业的深度融合与创新。“互联网＋”已经快速渗透到各个传统行业之中，而且不是简单的两两相加，而是利用信息技术以及互联网平台，让互联网与传统行业进行深度融合与创新，创造新的发展业态。例如，高等教育中的慕课热潮、在线教育、在线考试等“互联网＋”教育行动计划，都深刻地影响着学生对学习资源的掌握、查询、评价与利用及其社会实践训练的质量与效率。

1.5.2 创新创业能力培养的需要

信息检索素养是信息化社会人才素质的重要组成部分，一个信息检索素养良好的人，其判断力、决策力往往都较强。提高人的信息检索素养，是为了更好地开发与利用信息资源，是培养人们创新能力的基本需要，也是学习、实践、创新创业的基础。学生利用国内外快速发展的教育信息化环境，通过检索信息、收集信息、处理信息、创造信息，实现对知识的探索和发现，这对创新人才的培养具有重要意义。

"大众创业,万众创新"不是简单的口号,而是一种可行的全新社会理念、政府政策导向和高等教育人才培养行动计划。在"大众创业,万众创新"的时代背景下,创业创新日益成为综合国力竞争的制高点,而大学生作为最具创业活力和潜力的群体,如何培养其创业创新能力,是摆在当前社会发展面前重要而紧迫的任务。面对我国经济新常态背景下的经济转型升级,建设创新型国家和创新型组织,培养创新创业型人才,已经成为经济、社会和教育教学发展的基本理念。高等学校是现阶段开展创业创新教育的主阵地。创新创业教育是培养创新型与创业型人才的重要方式,是大学生树立创新创业意识、践行创新创业精神、形成创新创业能力的重要途径。

2015 年国务院"国发 23 号文"明确:大众创业、万众创新是富民之道、强国之路,必须着力创立大众创业、万众创新的新引擎。同时国务院办公厅《关于深化高等学校创业创新教育改革的实施意见》,进一步明确了高校作为青年创业创新人才培养摇篮的责任担当,深化高等学校创业创新教育改革,是国家实施创新驱动发展战略,促进经济发展提质增效升级的迫切需要,是推进高等教育综合改革,促进高校毕业生更高质量创业就业的重要举措。高校要厚植大众创业、万众创新土壤,为建设创新型国家提供源源不断的人才智力支撑。

培养学生的创新创业能力,就必须让学生主动地思考问题,独立自主地进行研究、探索、讨论、交流,在这种全新、宽松的学习氛围和环境中,学生必须具备较高的信息检索素养。信息检索素养较高的学生,能增加自我学习生涯规划和创新创业行动的机会,并在问题的独立思考、信息的选择与评估、信息的利用与反馈、创新创业实践的行动与总结过程中,不断提高学习效率与质量,提升创新创业精神与能力。

1.5.3　掌握有效信息和开展科研与学术活动的需要

科研与学术工作具有继承和创造的双重基因,科学研究的双重基因特性要求科研人员在探索未知或从事研究工作之前,应尽可能地占有与自身研究项目相关的大量信息,信息检索或信息查询是科学研究必不可少的前期工作,而良好的信息检索素养则是开展科学研究的有利条件。一项数据表明:一个科研人员的工作投入会有 50%用于查阅文献资料,32%用于研究,9.3%用于写研究报告和学术论文,7.7%用于思考问题,也就是说科研人员有一半以上的时间是参与信息交互活动。查阅各种书刊文献资料和网络数据库是科学研究的重要前提,凡是从事科学研究的人,在研究每一个新的课题时,仅在查找文献资料和相关数据库上所花费的时间,就要占研究课题总时间的 1/3,如果别人已经为你把这 1/3 花费了,这就会使科研人员加快课题的研究进程。如果别人已经查到某人正在研究

此课题,也许你就不会白白地浪费时间去做重复或者无用的劳动。人类的信息、知识、情报和文献资源每天都在急速增长。由于信息资源数量的急速增长,质与量都在不断地变化,任何一个科研工作者要想在茫茫的信息资源海洋中找到自己最需要的信息,如果不具备良好的信息检索素养,就只能是望洋兴叹,束手无策。如果在大学时期具备了良好的信息素养,就能利用信息检索原理、工具、方法与技术,充分了解国内外、前人和他人对你探索和研究的科研问题已做过了哪些工作?取得了什么成就?发展动向如何?这样才能做到心中有数,防止重复研究与资源浪费,将有限的时间和精力用于创造性研究。课题选题、立项阶段进行有效的信息检索,有助于理清思路,获得正确选题依据,提出质量高、内容新、有针对性的研究课题。只有这样,才能把别人的终点当成自己的起点,防止重复劳动,少走弯路。

从课题确立到整个科研研究过程,以及科研结束的成果鉴定等整个过程都离不开信息检索活动的支持和信息检索素养基因的能量传递,以判定成果的先进性、科学性和实用性。可以说,信息检索活动在整个科研过程中占有重要的位置。信息检索贯穿了科研工作的始终,是科研工作的重要组成部分。因此,大学生具备良好的信息素养,有助于在校期间的研究性与批判性学习,也有助于在校期间积极参与各类大学生科研项目、专业大赛和创新创业项目(例如各级各类大创项目、各种课程创新实践项目、各级电子设计大赛项目、各级大学生计算机设计大赛项目等),也更加有利于就业后积极参与工作单位的各种科研项目。

当大学生离开学校从事工作单位的科研工作时,良好的信息检索素养将有助于他们及时了解国内外最新的专业研究动态和科研成果,与国内外专家学者及同行进行交流,合理地制订自己的研究计划和科研进程,与工作单位的实际攻关项目与科研任务紧密结合起来,既可以少走弯路、快出成果,又能避免不必要的重复研究,多出原创性科研成果。

1.5.4 提供科学方法与正确决策的需要

科技、经济、学习、数字化生活等领域的管理与决策,同样离不开信息检索素养的支持。任何个人、企业,乃至国家,要想在竞争中立足,都必须掌握足够可靠的信息,并利用它进行科学决策,才能在竞争中取胜。如果要在浩如烟海的信息资料中盲目地寻找自己需要的信息,自然是一件困难的事情。管理决策必须依赖信息检索获取准确而全面的信息,才能保证其科学性、公正性与正确性。信息获取成功的基础则是通过科学合理的信息检索获取大量有用的信息。在激烈的市场竞争中,无论是企业还是国家之间,都时刻关注竞争对手的动向,力求扬长避短,确立自己的竞争优势,这种竞争的根基就是信息的竞

争。“优胜劣汰,适者生存”是市场竞争的自然规律。商场如战场,一个国家、一个机构、一个企业要想在激烈的市场竞争中立于不败之地,首先是要有科学的决策,信息竞争是进行科学决策的重要依据。企业在市场中要不断开发新产品,选择投资项目,确定营销策略,这一切都离不开准确及时的竞争信息。因此,信息竞争是企业成败的前提和基础,是企业决策的智囊、市场导向的风向标、市场投资的指示灯,是现代企业生存发展的战略武器和重要保障。

起源于国外先进且大型的检索工具与检索数据库,例如美国的科学引文索引(SCI)、工程索引(EI)和英国的科学文摘(SA)与世界专利索引(WPI)等,无不为科学决策提供了有力的信息检索支撑。

1.5.5　终身学习的需要

全社会已普遍形成这样的共识:唯有终身学习,才能培养完善的人;只有具备信息检索素养的人,才能实现终身学习,成为信息时代所需要的学习型与创新创业型人才。

终身学习是信息社会对人才教育与个人发展的基本要求。首先,知识本身具有发展进步性。大学生在校期间所学的知识会很快老化而失效,况且目前知识老化的速度又在日趋加快。与此同时,科学技术转化为生产力的速度却在日趋加快,从发明创造到应用推广的周期大大缩短。因此,以学历教育为目标的高等教育不再是各国高等学校的中心任务。其中,信息检索素养是大学生离开学校走向社会后,得以继续发展和进行终身学习的一项基本功。

大学生信息检索素养教育的内容之一,就是教会学生掌握知识,了解信息的组织机理;教会学生如何积累学习资料与学习资源,如何利用各种文献与数据库工具,如何利用现代信息技术搜索、查询、组织各种电子资源和网络信息;教会学生如何评价、管理和利用信息,使学生具有独立学习和终身学习所必备的技能和素质。因此,信息检索素养教育是人们终身学习的基本需要。

本章小结

在信息时代的今天,面对几何级数增长的海量信息资源,如何有效地检索、获取、评估、传播、共享和利用信息,成为了每个人重要的基本素养和能力要求。作为信息时代的大学生,要重视信息检索素养的知识学习与能力培养。在校大学生信息检索素养的理论知识学习与能力形成,在很大程度上也影响着大学生今后的生存与发展(例如终身学习的

需要）。

信息检索素养既是个体查找、检索、分析信息的信息认识能力，也是个体整合、利用、处理、创造信息的信息应用能力。具体描述为：善于根据问题分析自身的信息需求（例如学习或工作需要），进而确定信息来源并使用有效的检索或查找方法，及时地获取需要的信息；善于整理信息、分析评价信息，善于运用信息技术处理信息并用于解决问题；在信息的获取、处理、共享、使用的过程中有良好的信息意识、信息道德和强烈的社会责任心，有一定创新、协作和服务精神。

信息检索素养包含了检索技术和人文精神两个层面的意义：在检索技术层面上，信息检索素养反映的是人们利用信息检索的意识和能力；在人文层面上，信息检索素养反映了人们利用信息时表现出来的品质和修养（例如信息生成与利用的产权意识、信息安全保护意识、不良信息过滤与免疫、网络暴力抵制与防护、杜绝抄袭与剽窃等不端行为）。大学生要想在信息社会中更好地生存和发展，不断提高自身的学习、工作和生活效率就必须具备良好的信息检索素养。获取、评价、利用信息资源的知识和能力，已经在大学生的学习与研究、生活与娱乐、实践与工作等环节发挥着越来越重要的作用。

大学生信息检索素养的内涵可以包括以下几个方面。

（1）信息检索意识。信息检索素养教育最重要的一点是培养大学生的信息检索意识，即要求大学生具有一种使用计算机与其他信息技术来解决自己学习、工作和生活中信息需求问题的意识。

（2）信息检索伦理修养。大学生能够遵循信息应用的伦理道德规范，不从事非法活动，同时也知道如何防止计算机病毒和其他计算机犯罪活动，在法律法规允许的范围内合理合法的检索与利用信息资源。

（3）信息检索技术知识。掌握信息检索技术的原理、名词术语与基本应用，了解信息检索技术发展与作用，具有一定的信息检索技术知识，把握信息检索技术的发展与应用。

（4）具有一定的信息检索能力。即查询、评价和利用信息以提高学习、工作和生活效率的能力。能利用信息技术，获取自己所需要的信息，评价和分析所得到的信息，并有效地利用于自身的学习、工作和生活中。

信息检索素养具有普遍性、层次性、实践性与操作性、发展性等显著特征。信息检索素养主要包括信息检索意识、信息检索能力、信息检索道德、信息知识、信息观念、信息心理等方面，而信息检索意识（想到没想到，信息需求的有效获取意识）、信息检索能力（相应技能的会不会）、信息检索道德（信息共享与利用的合规合法性）又是其中的主要方面。

我国当代大学生的信息检索素养现状是：信息检索意识较弱、获取信息的检索能力不强、加工与利用信息的能力较差、信息道德和信息法规意识急需培养。大学生信息检索素养教育与培养的意义主要包括信息化社会对大学生的素质需求、创新创业能力培养的需要、掌握有效信息和开展科研与学术活动的需要、提供科学方法与正确决策的需要、终身学习的需要。

本章思考与练习题

1. 举例说明信息检索素养含义。
2. 举例并用“网络截图”说明你使用某一查询工具的方法与检索结果。
3. 美国图书馆协会阐明的学生具备信息检索素养的基本要求包括哪些方面的内容?
4. 信息检索素养具有哪些明显特点?
5. 在信息查询、获取、共享与利用过程中,如何做一个“合格网民”?
6. 信息检索素养主要包括哪些内容?
7. 大学生信息检索素养是否有评价标准,可以从哪些方面去评价?
8. 简述我国当代大学生的信息检索素养状况。
9. 大学生信息检索素养教育与培养有哪些重要意义?
10. 查询三篇关于“大学生信息检索素养教育”方面的学术文章,写 300 字左右的体会。
11. 作为网络时代的大学生,你认为应该具备什么样的信息检索意识?
12. 作为新时代的大学生,你认为应该具备什么样的信息检索道德品质?

第2章　信息检索与知识产权

信息社会是一个信息发现、信息挖掘、检索与利用、传播与分享的信息化社会，人们有可能充分发掘和自由利用社会共有的海量信息资源，并使之成为生产发展、生活质量提高和开展终身学习过程的核心要素。也正是基于信息的人类共享性和日益发达的信息技术手段，社会信息资源的保密、保护和专用及其个人信息安全受到了更严峻的挑战。日益发达的信息技术手段和无处不在的网络化环境，已经可以让人们轻而易举而又不露痕迹地检索、获取、共享和利用各种信息资源或信息产品，这就为各种信息化犯罪创造了条件；人们的信息检索与利用活动迫切需要形成更加广泛和深入的知识产权法律意识，也是信息检索道德的内在要求。

由于信息是知识产权活动的一种客观反映形式，而当代大学生作为信息社会信息检索与利用过程中最具活力的生力军，需要具备较高的知识产权法律意识，尊重知识产权，杜绝、避免知识产权侵害和各种网络化信息犯罪发生，共建公平、开放、和谐与守法的信息化与网络化环境。这不仅是顺利和合法开展信息检索与利用活动的前提与根本保证，也是当代大学生信息检索素养教育的内在要求。

2.1　信息与知识产权

2.1.1　信息

1. 信息社会

1）信息社会的概念及特点

信息社会又称信息化社会，也称信息时代。它是与工业化社会相对应的一种称谓，是一种以信息为标志，以信息技术为基础，以信息产业经济为支柱的社会。信息社会的主要特点有以下四个方面：

（1）在信息社会里起决定作用的不是资本而是信息和知识，信息成为比材料或能源更重要的资源。

（2）价值的增长不仅通过劳动，更重要的是通过信息与知识的掌握与创新。

(3) 人们更加关心和注重发展性需要,因而预测、检索、评价、传播与利用信息的重要性凸显。

(4) 以信息价值的生产、获取、传递、服务与利用为中心的信息产业经济快速成长并日益强盛。

2) 信息社会的三个核心要素

(1) 信息技术带动高新技术发展

本世纪以来,人类在互联网、新材料、新能源、生物、空间、海洋、航空航天等高技术领域取得重大突破和快速进展,其中信息技术的发展最为迅速。以信息技术为先导引发的高新技术崛起,构成了当代高科技发展的主流;而且信息技术及其成果向社会各个领域的渗透和广泛利用(例如移动互联网、"互联网+"等信息技术),促进了高新技术的深度开发与交叉融合,也为知识产权法律机制运行提供了更广阔的技术基础。

(2) 信息产业促进传统产业结构加速调整

包括信息设备制造业与信息服务业在内的信息产业的飞速发展,不仅大大加强并迅速提高了第三产业的质和量,而且促进了第一产业和第二产业的深度调整。信息产业实现了社会产业结构的再调整与革新,其中信息服务业发展水平则成为一个国家、一个地区或者一个行业发展程度的重要标志。而市场经济中信息产业和信息经济的发展,更需要包括知识产权法律在内的国家政策和法律约束来引导和规范全社会的信息活动与个体的各种信息行为。

(3) 信息资源引导经济集约化

经济强国主要利用全世界范围内的广泛信息资源形成经济集约发展,而我国经济发展相对比较粗放的重要原因之一就是信息具有资源替代功能。信息技术的发展为社会信息资源的开发与利用提供了高效、便利、平民化的普惠条件。在信息资源数量急剧膨胀的网络化环境中,信息资源的管理与利用就显得尤为重要,它成为经济集约化发展的一个关键因素所在。整个信息的检索与利用过程都应当遵守的规则之一,就是尊重知识产权、遵循知识产权法律,也就是信息的检索与利用活动必须在依规合法的前提下进行。

2. 信息的含义

"信息"一词在英文、法文、德文、西班牙文中均是"information",日文中为"情报",我国台湾称为"资讯",我国古代用的是"消息"。作为科学术语最早出现在哈特莱(R. V. Hartley)于1928年撰写的《信息传输》一文中。信息,通俗地称为音讯、消息;通信系统传输和处理的对象,泛指人类社会传播的一切内容。人通过获得、识别自然界和社会的不同

信息来区别不同事物，得以认识和改造世界。在一切通信和控制系统中，信息是一种普遍联系的形式。1948 年，数学家香农在题为《通信的数学理论》的论文中指出："信息是用来消除随机不定性的因素"，这一定义被人们看做是经典性定义并加以引用。美国数学家、控制论的奠基人诺伯特·维纳在他的《控制论——动物和机器中的通信与控制问题》中认为，信息是我们在适应外部世界，控制外部世界的过程中同外部世界交换的内容的名称。

信息是事物运动的一种状态与方式，是物质的一种属性。

信息不同于消息，消息只是信息的外壳，信息则是消息的内核；信息不同于信号，信号是信息的载体，信息则是信号所载荷的内容；信息不同于数据，数据是记录信息的一种形式，同样信息也可以用文字或图片来表述。

总之，"信息是事物运动的状态与方式"这个定义具有较大的普遍性，它不仅能涵盖所有其他的信息定义，还可以通过引入约束条件转换为其他的信息定义。例如，引入知识主体这一约束条件，可以转化为认识论上的信息定义，即信息是认识主体所感知或所表述的事物的运动状态与方式。层层引入的约束条件越多，信息的内涵就越丰富，适用范围也越来越小，由此构成相互间有一定联系的信息概念体系。

3. 信息的类型

信息的类型可以从不同角度来认识。了解信息的类型有助于人们加深对信息内涵及其特征的认识，丰富信息检索与利用的知识。

（1）按产生的客体区分。从产生信息的客体的性质来分，信息可分为自然信息、生物信息、机器信息和社会信息。

（2）按存在的形态区分。信息的形态可分为媒介形态和符号形态。信息的媒介形态以其所依附的载体为依据，可分为文献信息、声音信息、电子信息等；信息的符号形态是指用于指代客观事物的字母、电码、语言符号等象征物，可进一步分为语言符号和非语言符号。语言符号是信息传播的主要象征，是人与人之间进行交流的工具，如作为汉语书面符号的文字、汉语拼音和汉语速记。非语言符号在人际传播中具有表露情感、替代自然语言、辅佐语义表达和调节行为的作用。

（3）信息分类还有其他的划分方法。如以信息的记录符号为依据，可分为语音信息、图像信息、文字信息、视频信息、音频信息等；以信息的运动状态为依据，可分为连续信息、离散信息；以信息的加工层次而论，可分为初始信息和再生信息，后者是对初始信息进行分析、加工处理后的结果，有时称为知识信息，也是信息检索的主要对象（例如图书、期刊、论文、报告、专利、标准、手册、指南、电子数据库、网络数据库等）。

4. 信息的特点

(1) 无穷性。信息是物质存在的一种方式或活动状态,而物质处于无穷的运动之中,这就决定了信息的无限性。物质和能量是无穷的,信息同样也取之不尽,用之不竭,人类将依赖这三大资源生存和不断发展。

(2) 可辨识性。信息可以通过人的各种感官直接辨认,也可以用各种技术手段间接识别。因为信息的载体形式是多样化的,因此不同的信息可以用不同的方法进行辨认。

(3) 可转换性。信息可以从一种载体形式转换为另一种形式,如物质信息可转换为语言、文字、图形、记号、代码、信号等。每个信息载体之间又可互相转换,可以从语言转化为其他代码,从图形转化为文字,从纸质载体转换为网络数据库等。

(4) 可存储性。人类可以用大脑将信息存储为"内语言",这是一种隐性存储形式,也可用机器设备存储信息,如纸张、光盘、磁盘、网络服务器等。

(5) 可扩充与紧缩性。事物不断运动,信息不断弃旧更新,社会的信息总量在不断增添与扩充过程中,而经由人们对信息的收集、加工、检索、传播、概括、融合、应用、再创造,又可以将信息容量大大收缩,以利于进一步发挥信息的潜能。

(6) 可替代性。准确而高效使用信息,可以进一步发挥其效力,减少各种社会工作、学习和生活的实际耗费,例如资金、智力、体力、物质和能源等实际消耗得以显著降低,因此在很大程度上,信息的可替代性特征十分显著。

(7) 可传递性。这是信息的重要本质属性之一,信息的功能与作用是通过传递特性实现的。信息也只有通过传递,才能发挥其"消除事物不确定性"的功能。

(8) 可分享性。可分享性也称为共享性,信息可以被分享,除了一些特定的信息和一些特定的人群外,在一定规模内被传递出来的信息一般是这个范畴内的每一个个体都可以分享的。

(9) 可组合性。若干信息被人有意无意地组合或融合起来,就会形成与本来信息不同的新信息。例如蒙太奇就是一种信息的组合,蒙太奇来自于法语 Montage,原意为构成、装配,是指"在影视作品的创作中将一个个镜头,依据一定的逻辑关系任意组接在一起,以表达需要的视频意义"。这一个个的分镜头就是一个个分支信息,进行不同的组合能发生不同的意义和结果。

(10) 非完整性。任何信息都不可能,也不必要反映出客观对象的各个方面,它只是事物的某一方面的某一种变化的反映和变化。

2.1.2　知识产权

1. 知识产权的含义

知识产权主要是指人们对其从事智力活动而产生的成果所依法享有的专有权利，是一种无形财产权。知识产权是人类的发明创造、智力活动成果和法律活动的结合与交叉，是人们依据国家法律对自己的智力活动而获得的成果所享有的权利。知识产权是一种看不见、摸不着的无形财产权，它能通过使用和有偿转让等多种形式创造财富，让它的拥有者从中受益。世界贸易组织在《与贸易有关的知识》中规定，知识产权包括：①著作权和邻接权；②商标；③地理标志；④工业品外观设计；⑤专利；⑥集成电路布图设计；⑦未公开的信息。

2. 知识产权的主要特点

（1）专有性。它又可称独占性、排他性、垄断性，是其权利人所依法拥有的专有权利，他人不得侵犯。

（2）地域性。它是指国家确认和保护的知识产权，只在该国的地域范围内有效，对其他国家不发生法律效力。

（3）时限性。知识产权一般只在法律规定的期限内有效。过了有效期，相应的智力劳动成果就成为任何人都能合法使用的社会公共财富。

（4）无形性。知识产权与有形财产不同，没有具体的形体。尽管知识产权需要依附于有形载体而存在，但无论是智力创造或知识成果本身，还是附载于工商标志的信誉都是“无形”的。

（5）法定性。知识产权的取得一般要履行相应的行政审批程序（例如各种类型专利），但著作权和商业秘密除外。

3. 知识产权的性质

知识产权是一种民事权利。它所反映和调整的社会关系是平等主体的公民、法人之间的财产关系，从而具备了民事权利的本质特征。知识产权的发生、行使和保护应适用民法的基本原则和民事规范，如民事主体、客体、内容、法律事实、民事法律行为等。

知识产权同其他民事权利一样，是一种私权。私权是与公权相对应的一个概念，指的是私人（包括自然人和法人）享有的各种民事权利。知识产权具有私人财产权利的基本特性。

知识产权是一种不同于财产所有权的无形产权。世界知识产权组织认为：知识产权与有形财产的最主要的不同点在于，对于诸如一张桌子，所有人可以通过占有它而基本达到保护自己的财产不受侵害的目的；而对于诸如一项发明、一部作品或一个商标，所有人

基本上不能通过占有它们而达到保护它们不受侵害的目的。

知识产权客体的非物质性是知识产权区别于财产所有权的本质特性。知识产权的客体即一定的信息内容，是没有形体的、非物质性的。客体的非物质性是知识产权的本质属性所在。当我们买卖有形商品时，转让的是该有形商品的财产所有权，而财产所有权的客体就是该有形商品本身，我们可以通过占有来实现转让。而转让知识产权时，转让的是知识产权本身，而不是载有信息的有形载体的财产所有权，载体的转移并不等于知识产权的转移，知识产权的转让也无须载体的转移。知识产权的客体是非物质性的有关信息（例如专利领域中的技术方案、著作权领域中的作品、网络原创性视音频信息等）。作为财产所有权客体的物体，是可以被特定人占有的，而作为知识产权客体的信息（如技术方案、商标标识或作品等），则不可能被特定人占有（它们可能被无限地查询、复制、传播和分享），因此可能被无限数量的人占有。例如，某人在其购得的一张光盘中刻录了某个计算机应用软件，他通过合法占有这张光盘而成为财产所有权人，但绝不会因其再次或多次刻录（复制）该软件，就可以“占有”在其光盘中的软件的知识产权。

4. 信息对于知识产权的意义表现

（1）信息是知识产权活动的一种反映。知识产权是一种人类法律活动，作为一种客观事物，它是人类社会的一种客观存在现象。信息既是这种社会客观存在的表现形式，也是对这种客观事物的反映。

（2）信息是知识产权现象的表述。知识产权现象不能够自我显示和表述，信息在知识产权发生与发展中同时产生，信息所要表示的目的是表达和显现知识产权作为客观事物的存在，因而这种信息是知识产权的重要属性之一。

（3）信息是人们认识知识产权的中介。人们在从事政治、经济、技术等活动中都要接触和利用知识产权，而人们认识知识产权现象则必须通过显示知识产权存在方式的信息，因而人类活动离不开信息，信息是沟通认识主体和认识客体的中介和桥梁，是人们认识和利用知识产权的必要途径。

2.1.3　知识产权信息

1. 知识产权信息的含义

知识产权信息是表征知识产权属性的信息，这种属性既包括知识产权作为整体的属性，又包括知识产权内各种具体智力成果与权利的属性，同时知识产权信息又是表达知识产权保护客体内含的信息，它包括有专利信息、商标信息、著作权信息、技术合同信息、涉及知识产权业务的竞争信息等。因而，知识产权信息概念可以包含两层含义。

（1）知识产权保护客体的内含信息。专利文献、商标文献、著作权作品中所包含的信息以及工业产权与著作权开发、交流、传播中的信息，都是这种客体内含信息。

（2）知识产权的“成长性”信息。这种信息主要是指知识产权的产生、发展、变更中所发生的信息。

2. 知识产权信息的一般结构

（1）信息内容。指由知识产权信息所表述的各种知识产权及其客体的内容。

（2）信息载体。知识产权信息载体既有物质材料载体，又有人工载体，还包括实物载体、大众媒体等。

（3）信息符号。一切信息符号都可作为知识产权信息符号。例如文字、图形、代码、音频、视频、语言等。

3. 知识产权信息的本质特性

根据上述对信息的一般描述和信息对知识产权的意义，我们可以这样来描述知识产权信息的本质：知识产权信息是知识产权存在方式和存在状态的表述和反映，知识产权信息是知识产权的重要属性之一，是显示知识产权存在的一种特性，知识产权信息又是人们认识和利用知识产权的中介。

知识产权信息具有信息的一般特征：①普遍性，即知识产权信息广泛存在于知识产权的各个环节之中。②无限性，即知识产权信息可以再生，可以不断地开发与利用。③特殊商品性，知识产权信息不仅具有价值和使用价值，而且其价值在通过交换过程实现后却并不失去使用价值。④载体性，即知识产权信息总是依附于一定的载体而存在。⑤共享性，即知识产权信息可以被众多的使用者所共享。⑥可伸缩性，即知识产权信息可以根据需要加以控制，可扩大或缩小而内容不变等。

2.1.4 知识产权信息的概念特征

在法律上和在现实社会经济活动中，知识产权都是一个抽象性、概括性的概念，它是对专利权、商标权、著作权（包括软件著作权等）以及后来扩展到的各种智力成果权（甚至延及反不正当竞争行为）的一种宏观概括和哲学意义的升华。知识产权信息也具有这种概念特征，在法律上人们具体引用的、在经济交往与文化交流中人们具体涉及的，都是各种具体知识产权信息，如专利信息、商标信息、著作权信息等，因而，知识产权信息是对这些具体知识产权的哲学概括和升华；专利信息、商标信息等为知识产权信息的总结提供了条件、素材和具体内容，知识产权信息则是对专利权、商标权、著作权等发生与发展过程中发生的信息的抽象规定。它不仅代表着人类有关知识产权信息的概念产生，而且有利于

人们确认知识产权的客观存在方式和状态，并且知识产权信息也是对知识产权结构及内容体系的自我完善。

2.1.5　知识产权信息的内容

知识产权信息是关于知识产权保护客体内含的信息，同时又是有关知识产权权利的信息，因而它有着十分丰富的内容。

1. 人类认识信息

知识产权保护客体涉及人类科学技术、文学、艺术、商业活动领域，是有关人类在这些领域从事智力劳动所创造的认识成果，因而知识产权信息首先是人类有关科技、文学、艺术、商业活动的认识信息。

2. 法律保护信息

知识产权信息基于法律活动而存在，因而它必然表现与显示法律活动的存在状态。

3. 知识产权贸易信息

（1）知识产权贸易主体信息。

（2）知识产权贸易标的信息。

（3）知识产权贸易方式信息。

（4）知识产权经营规则信息。

（5）知识产权价值计量信息等。

4. 智力成果的形象信息

知识产权客体往往借助于具体事物，形象地表达设计、开发与创造思想，因而知识产权信息是一种丰富的形象信息。

5. 信息知识产权的各种载体

知识产权信息具有流动性，它既存在于法律管理部门的内部保存或面向社会公开的各类文件之中，又存在于缩微文档、机读数据库和互联网之中，同时还存在于实物商品、市场销售之中，因而知识产权信息载体形式多样，具体包括的各种载体有：①印刷型文献；②缩微文献；③机读电子文献；④网络数据库。

2.2　信息检索与利用的法律规范和信息道德

不同的信息主体，如国家、组织和个人，享有不同的信息权利，也承担着不同的义务。信息时代的数字化技术和网络化发展，在一定程度上改变了人们的价值观和伦理观，驱使

某些信息需求者为达到目的而不惜动用一切技术手段,包括非法的或介于合法与非法之间“打擦边球”的方式去发现、评估、检索、获取、占有和使用信息资源。

作为当代大学生,通过各种信息检索技术手段以满足自身不断增长的信息需求,是时代的要求与必然趋势,但前提是不能侵害他人或组织的正当知识产权利益,不得干扰或危害和谐的信息资源利用环境和信息共享秩序。在获取与利用信息的同时,信息检索与利用者的检索手段和利用方法也必须置于法律和道德允许的范围内。

2.2.1 信息检索与利用的相关法律制度

1. 信息检索与利用的法律属性

1）信息产权的法律属性

信息产权以能被人们认识、感知和了解的信息为客体,具有价值财产,并能给所有者或权利人带来经济利益,能在市场交换中给信息拥有者或信息权利人带来物质和精神财富,可以成为产权交易的对象进入资本市场;信息产权在内容与形式方面具备法律规定的要件,所有人或其他信息权利人在信息查询与搜索、使用与转让、加工与存储、复制与修改等活动中享有人身权与财产权。其财产权属性主要体现在以下两个方面。

第一,信息产权是信息产权人或其他信息权利人直接控制、支配其相关信息并排除他人非法侵害的权利,是权利人就相关信息的查询、存储、处理、加工和传播等过程中合法使用、利用而获取利益的权利。

第二,在市场交换条件下,相关信息能作为信息权利人与市场商业主体交易的客体,其财产权属性主要体现在大量的信息再加工、深度开发而形成的数据库产业所具有的财产权属性;权利人能够通过对信息使用权的转让直接获取财产利益。

2）信息检索与利用活动的法律属性

信息检索是信息使用者根据自己的需要,利用有关信息检索知识,通过各种检索途径和检索工具获取相关信息的过程。信息检索活动具有明显的目的性,检索过程中可能侵害他人的利益。检索对象受到著作权法、专利法、商业秘密保护法、民法通则、劳动法、保守国家秘密法、刑法等法律法规的保护。因此信息检索与利用活动在法律上表现出自身的本质和特征,具有明显的法律属性。

目前,世界上没有专门的信息检索与利用法规,其依据依托在相关知识产权法领域。但是,无论在理论上还是在实践中,其法律属性都是无法回避的问题。由于信息检索与利用对象受到各种法律法规、行为准则的限制,信息检索与利用活动在很大范围内涉及法律规范的问题。

2. 信息公开制度

所谓信息公开，一是指政府有义务公开自己的活动情况，包括行使管理职责过程中形成的各种信息；二是指公民个人或团体有知情权去了解、查询、获取和应用行政机关的文件、档案资料和其他信息。

建立信息公开制度对于信息检索与利用的意义。首先，每个合法公民都有信息检索与获取的自由和权利，建立信息公开制度可以增加政府活动的透明度和公开性，方便民众对信息活动的查询与获取。其次，建立信息公开制度，便于实现政府上下级之间、各部门之间的信息查询与利用，避免信息堆积和重叠，构建高效率的信息化政务。最后，有利于社会信息最大程度地共享。在信息化社会，信息公开对于公众及时获取所需信息，降低其查询与利用成本并促进整个社会经济的发展具有重要意义。

2.2.2　知情权问题

知情权就是为了公民和企事业组织、社会团体了解与自身利益紧密相关的资料、信息、消息而建立的保障制度。在我国，除了司法与公安人员有依据法律规定进行调查、取证的权利外，公民也有权通过正当的途径查询和搜集法律允许获取的信息，如《中华人民共和国行政复议法》当中规定的利益相关人的查询制度、行政处罚法规定的听证制度等。随着社会的不断进步，知情权成为一种广泛的社会权利和公民权利。

1. 知情权的含义

“知情权”，英文名为 the right to know，又称为“知”的权利、知悉权、咨询权、信息权或了解权，是《世界人权宣言》确定的基本人权之一，其基本含义是公民有权知道他应该知道的事情，国家应最大限度地确认和保障这一权利。

“知情权”包括“知悉”、“获取”两个层次的含义。其中“知悉”主要是指权利人从主观上知晓，而“获取”则指权利人查询、索取、查阅某种记录着信息的有形载体（这种载体可以是文字、图片，也可以是录音带、录像带、电子光盘或网络数据库等）。知情权表达了现代社会成员对信息资源的一种普遍利益要求和权利意识，是一个民主社会和民主国家中公民享有的重要权利。

2. 中国现行法律中对“知情权”的规定

我国已陆续制定了一些法律规范，以明确和保障公民知悉、获取有关信息的自由和权利。像《宪法》、《消费者权益保护法》、《证券法》、《公司法》、《合同法》、《保险法》都列出了相关条例保护公民的合法知情权。此外，还有属于保障公民获知信息的义务性规范，以及公民有权通过各种渠道享受国家机关和其他公共团体依法提供的信息服务。

2.2.3 国家秘密问题

1. 国家秘密的概念及其构成要素

《中华人民共和国保守国家秘密法》规定，国家秘密是指“关系国家的安全和利益，依照法定程序确定，在一定时间内只限一定范围的人员知悉的事项”。这一概念表明，一条秘密信息必须同时具备以下三个要素才能算是国家秘密。

首先，关乎国家的安全和秘密。这是构成国家秘密的实质要素，是准确判定某一信息是否属于国家秘密的关键。

其次，必须根据国家相关法规、依照法定程序加以确定，这是国家秘密的程序要素。强调确定国家秘密的统一性与合法性，防止主观随意性。例如，高校少部分优秀学生可能参与的保密性科研项目(例如国防正式或预研项目)，其研究成果是不允许公开发表的，与涉密项目有一定关联性的成果，确实需要公开发表的，也需要依规合法进行相关审批后再发表。

最后，具有特定的保密时限和限定的知密范围。这是国家机密的时空要素，是为保守某项国家机密，需在一定范围内所采取的各种保密措施。

2. 国家秘密的密级

《中华人民共和国保守国家秘密法》第九条规定了国家秘密的基本范围，主要内容有国家事务重大决策中、国防建设和武装力量活动中、外交与外事活动中的秘密事项，以及对外承担的保密义务、国民经济和社会发展中、科学技术研究中、维护国家安全活动和追查刑事犯罪中的保密事项及经国家保密行政管理部门确定的保密事项。

根据秘密等级，国家秘密可分为绝密、机密和秘密三种。识别国家机密，主要通过信息载体上的国家秘密标志，即印记在国家秘密载体上，表明其内容属于国家秘密事项的记号。一个完整的国家秘密标识应为国家秘密的密级、五角星符号、保密期限。

3. 有关法律对国家秘密的保护

国家秘密，其法律保护是以刑法和行政法为主的多种法律保护体系，主要包括《中华人民共和国宪法》、《中华人民共和国刑法》、《中华人民共和国国家安全法》和《中华人民共和国保守国家秘密法》。

在信息检索、收集、获取与应用活动中，任何人均不得非法查阅与使用国家秘密。一旦在工作中发现有国家秘密材料，应立即联系政府保密部门，并妥善保管这些材料，移交政府保密部门。最后，还应尽最大努力协助调查这些资料是如何泄露出来的，同时也需要妥善保管自己信息检索与获取过程中的相关资料，以备政府保密部门调查。

2.2.4　商业秘密问题

1. 商业秘密的含义

商业秘密，是指不为公众所知悉，能为权利人带来经济利益、具有实用性并经权利人采取保密措施的设计资料、开发程序、产品配方、制作工艺、制作方法、管理诀窍、客户名单、货源情报、产销策略等技术信息和经营信息。

2. 有关法律对商业秘密的保护

目前，很多国家和地区采取不同的方法和手段保护商业秘密。除少数国家和地区通过专项立法、对商业秘密加以保护外，世界上大多数国家和地区都有通过《中华人民共和国民法》、《中华人民共和国反不正当竞争法》以及《中华人民共和国刑法》等有关条款予以保护。我国现在还没有针对商业秘密的保护问题专门立法，对商业秘密的保护主要是通过《中华人民共和国反不正当竞争法》、《中华人民共和国劳动合同法》、《中华人民共和国民法通则》和《中华人民共和国刑法》的有关规定来实施保护的。

3. 信息检索与利用过程中的商业秘密的合法获取与应用

在信息检索与利用过程中，不能侵犯他人的商业秘密，但是不等于不能利用一定的途径合法获取商业秘密，或与商业秘密价值相对应的信息，依规合法获取商业秘密在一定范围内是允许的。

2.2.5　隐私权保护问题

原则上属于个人的信息，都是个人秘密或隐私，都应得到保护。个人信息的隐私敏感度在不同的地域是不同的，由于社会习俗的不同，同一项信息的隐私敏感度在不同国家、不同地区、不同民族之间是有较大差异的。

1. 信息检索与利用过程中需避免的侵犯隐私权行为

信息检索与利用过程中，需要避免的具体侵害行为有：侵入侵扰、监听、监视、窥视、窃取、刺探与收买、搜查、干扰、披露、公开或宣扬。

2. 网络信息时代信息检索与利用过程中存在的侵犯隐私权问题

所谓网络隐私权，是指在网络环境下借助互联网而享有的个人生活安宁和私人信息不受他人信息侵害的权利。因此，通过网络环境进行信息检索与利用过程中，注意以下行为构成侵权：①在用户不知情的情况下的信息搜索、获取与应用；②电子邮件、通信软件和社交媒体的监视、篡改与冒名；③被采集信息或残留信息保护不当；④隐私权客体范围被恶意分享与传播扩大化。

2.2.6 信息复制权保护问题

1. 复制权的法律保护

复制是信息检索、评估后进行利用的一种重要方式，而复制权也是信息权人最重要的权利，通过对作品复制权的控制，信息权利人能充分行使自己的使用权。

我国的《中华人民共和国著作权法》第十条、《最高人民法院关于审理关于计算机网络著作权纠纷权案件适用法律若干问题的解释》第三条、《互联网出版管理暂行规定》第五条都做出了复制权保护的相关规定。

2. 信息检索与利用过程中存在的复制权问题

在英、美等国家的法律中，对版权的限制有一个原则：信息的合理使用，就是对发表的信息作品，可以不经著作权人的许可，不向其支付报酬就可以使用。通常情况下，批评、评论、教学、个人学习、学术或研究一般都包括在合理使用的范围内。因此，要对信息合理地采集、检索与应用，应考虑的因素包括信息查询的目的与特征，即该检索与利用活动是否具有商业性质，或者是否是为了非营利的教学与学习等目的。要对信息检索对象的性质、所获取与利用信息的质与量以及采集对象作为一个整体的关系来考虑，甚至要考虑信息检索与利用对象的潜在市场应用或学术价值所产生的影响。

在网络化社会的今天，作为大学生在学习和研究活动中，切忌不尊重他人的原创性信息或原创性知识成果，在实验报告、课程设计、社会实践、学术论文发表或毕业论文撰写等过程中坚决抵制剽窃、抄袭(有意图地复制与粘贴)、盗取、拼凑、伪造、篡改、买卖等各种学术造假与学术腐败的违法行为。

2.3 信息检索与利用过程中的道德自律

法律与道德之间的区别并非总是一清二楚的，存在模糊的灰色区域，在这些区域，有合法与不合法、道德与不道德。不同国家、不同法律约束范围、不同行业对道德的含义都有不同的定义与诠释。为便于理解，可以将其简化如下：法律是他律性约束，非法行为将承担民事或刑事后果；遵守信息道德是为履行所在专业、职业、行业及其他行业标准的行为。信息检索与利用行为不道德也同样损害自身的利益。因此，除了行业自律、乡规民约和舆论监督外，人们更多应靠建立自我约束的机制，即通过学习和教育提高信息道德意识和自觉性，达到自我管制、自我约束的目的。在信息化社会中，大学生的信息道德意识与信息道德自律的培养与形成，也是信息检索素养教育的一项主要内容。

2.3.1 法律约束的局限性

信息社会中,信息的海量增长、网络化关联、易检索性、易复制性、易扩散性等特点,决定了完全依靠刚性约束为主的法律手段来规范信息检索、获取与利用者的行为是不现实的,也是比较困难的。

以保护企业的自主知识产权、防止侵犯商业秘密的行为为例,法律约束的有限性体现在以下几个方面。

(1)《中华人民共和国著作权法》对未公开的自然科学、工程技术作品给予保护,但只保护其表现形式,不保护其内容。如果按内容去实施其中的技术秘密、经营秘密,则不构成侵权。

(2)《中华人民共和国专利法》的最大缺陷是取得专利必须将自己的技术做彻底公开,以换取此项发明在一段时间内的垄断权,这就为竞争者从公开的专利文献中分析有用信息开启了方便之门。

(3) 从法理上讲,商业秘密权不具有绝对的独占性和排他性,行使的仅仅是相对权利。其禁止效力仅涉及违法侵占,但不及于合法取得。

(4) 使用《中华人民共和国侵权行为法》保护商业秘密,有两个潜在的困难:一是受害人必须证明自己是某一个合法权利(或利益)的享有者,而信息所有权的认定在某些情况下是极为复杂的;二是受害人必须证明侵权人的主观过错。实际上,商业秘密是一种脆弱的权利,在许多情况下,雇主还来不及获取实际的侵权证据,其商业秘密就已经丧失了。

(5) 通过合同保护商业秘密的严重缺陷就是:合同的效力在通常情况下并不涉及合同当事人之外的第三者。

(6)《中华人民共和国反不正当竞争法》所反映的行为只是不正当竞争行为,而不是所有的竞争行为。

(7) 在现实的诉讼与纠纷中,商业秘密侵权的事实认定往往需要物质证据,如记录有秘密的文件或操盘等。

(8) 商业秘密禁止在实际操作中的合理尺度仍然难以把握。如果雇员和雇主事先并没有签订商业秘密禁止协议,或者协议无效及不完善,则仍无法阻止对雇主的潜在损害。

所以在信息社会中,信息道德的培养与形成是十分必要的。

2.3.2 信息道德自律问题的提出

遵守信息道德,是信息化时代和网络化环境下全部个体与组织的共同利益需求,信息

道德失范行为不符合他人的利益和公共利益，归根到底也有损于自身利益。

参与竞争的从业人员常见的观点是：对抗性的竞争不需要也不能讲伦理道德。所谓“法无禁则自由”，信息检索与利用活动只要不违法就可以，由此造成主观刻意所为的“钻空子”、“打法律擦边球”的行为比比皆是。

法律对社会的控制与调节是通过国家机器的强制手段实施的，属于他律范围。无论从时间还是空间方面，法律应用于复杂的社会关系时总有一定的局限性，更多的规范应依靠道德自律。道德作为一种行为规范，与强制性的法律规范不一样，它主要是通过两个方面的相互作用来实现规范和约束人的行为：一是社会舆论的评价与监督；二是行为主体的内心体验。对于某一种行为或者某单一行为的道德价值判断标准，是一个民族、一个国家长期政治、经济、文化、宗教等因素相互作用而形成的，具有一定的稳定性。

2.3.3 信息道德的培养和内省原则

（1）信息道德的习得与实践相结合。大学生的信息道德可以在教育过程中通过习得的方式逐步形成，同时在各种信息检索与利用的实践过程中逐步提高信息道德品质。“知识即美德”，把知识本身当做美德，在学习中成长信息道德品质；“省察克己”，在信息检索与利用过程中内省自身的信息资源利用目的与主观意图，是否侵权、是否涉密、是否侵害他人隐私等不道德或违法行为。

（2）信息道德的自律和他律相结合。不管是什么样的信息检索与利用活动首先需要自律，但自律不意味着不要规则、不要法纪，需要自律与他律相结合，从而保证开放、共享、和谐与繁荣的信息资源生态。

随着人们法制观念的逐步增强，“遵纪守法是合格公民的内在要求”已成为普遍共识。信息检索、获取与利用的道德操守，归根到底是人文精神体现，它关乎信息化社会的健康发展与繁荣。

总之，知识产权法律为信息检索、获取与利用画出了一道红线，也保证了信息检索与利用活动的效率与质量，但是为了避免进入误区，引起不必要的信息与知识产权纠纷，甚至侵害个人隐私或危害组织机密等不当行为，信息的检索、获取与利用还需要道德自律。

2.4 信息检索与利用同知识产权保护的相互影响

2.4.1 信息检索与利用对知识产权保护既制约又促进

在信息检索与信息资源共享过程中，知识产权保护的客体——信息和知识产品，始终

处于非物质状态，具有作为生产要素的需求性、共享性、易复制性和扩散性等特点。正是这些特性的存在，再加上信息技术及网络通信技术的日益发达，使得它们可以轻而易举地被利用者检索、获取、复制、传播和使用，而需要付出的成本很低或者几乎没有。因此，这种主要基于信息资源共享为目的的检索与利用活动就会容易产生侵权和违法行为，而且信息侵权与违法行为在网络技术逐步发达与不断渗透的当下，更容易泛滥成灾，进而破坏信息与知识产权生态，使得信息生产者不能回收自己在信息生产过程中所付出的智力成本和物资成本，挫伤了他们进一步进行信息生产与知识创造的积极性和主动性。

为了解决信息时代不断出现的诸多信息检索与共享失范行为和侵权行为，需要不断研究、细化、修订和丰富传统知识产权法的内容，进一步应用法律的手段来规范人们检索、共享、传播和利用信息资源的行为，保护信息生产者的利益。新的信息时代呼唤新的知识产权法，因此，信息资源的检索与共享虽然在一定程度上制约了知识产权保护的作用，却在无形之中又促进了知识产权制度的发展与完善。

2.4.2　知识产权保护对信息检索与信息资源共享的制约和促进

众所周知，信息和知识产品，其本身就具有公共物品属性。从经济学角度看，它们一经产生出来就应该进入公共领域，不受限制地为可能利用的人打开方便之门，供人们自由查询、获取和利用。然而，知识产权保护制度使得这种现象不复存在，它规定知识产权具有专有性：即知识产权归属权利人所有，他人如要使用该项智力成果，必须得到权利人的许可，并向其支付一定的报酬。知识产权的这种专有性和智力成果的有偿性决定了某些信息资源被权利人所垄断，不能为社会公众自由获取和利用，这使得可供共享的公开的信息资源数量减少，影响了公众对信息查询与获取的广度和深度，严重时会导致信息闭塞，使社会公众利益受损。由此可见，知识产权对信息资源的检索与共享是有制约作用的。

知识产权法的本质是平衡知识产权人和社会公众利益的调节器，“寻求私人利益与公众利益的平衡”一直是知识产权法追求的目标。也就是说，知识产权法在制定的时候就已经在尽力协调、平衡和兼顾权利人与社会公众的利益了。一方面，它要保护信息和知识生产者的利益，允许他们向利用者收取报酬来补偿自己的投入，并为之带来一定的经济利益，从而鼓励他们继续创造更多的信息和知识，也就是保护了信息获取与共享的“源泉”；另一方面，它又要兼顾社会公众的利益，防止权利的滥用和过度膨胀，促进信息和知识的广泛传播、共享和利用，加快社会进步。因此，知识产权保护对信息检索与信息资源共享又起着促进作用。

由此可见，信息资源检索、共享与利用同知识产权保护之间既存在着相互矛盾的一

面，也存在着相互统一的一面。如果没有合法的信息与知识保护就不会有源源不断的信息与知识产生，信息有了合法的保护就会有日益丰富的信息可供查询、共享与利用。

2.5　大学生信息检索素养与学术不端行为的关联

作为大学生，信息检索的主要目的是为了自主学习、发现与探究学习、协作与研究性学习、课题与项目研究活动及其生活与休闲娱乐等活动服务，在满足信息需求的基础上，提高学习、研究、工作和生活的效率与质量。但是，大学生侵犯他人知识产权、侵害别人智力劳动成果与学术不端行为却屡屡发生。

高校是人才培养和学术发展的主阵地，是传授知识、传播知识、利用与生产知识的圣洁天堂。然而近年来，来自于大学生的“科学骗局”、“困境中的科学”、“伪造的结果”等学术不端现象频发，严重影响了大学生作为高级知识分子和高层次人才的社会声誉，并且对大学生的学术能力成长、学术品质的形成与信息检索素养的塑造构成威胁，大学生的学术不端行为也呈现出一些腐化与泛化的不良状态。因此，对大学生进行信息检索道德教育、学术规范引导和学术道德培养，遏制其学术不端行为，优化大学校园的良好学风与学术风气，已势在必行。

2.5.1　大学生学术不端行为的界定

学术研究必定包含着诚信、客观、借鉴、参考、合作与创新等价值，现代科学进步是学术进步取得成功的重要体现，学术诚信是一个历久弥新的话题。早在 1989 年美国公共卫生署就将“不端行为”定义为：伪造、篡改、剽窃或在研究的申请、执行或报告过程中严重偏离科学界公认的科研行为准则的行为，但不包括无意的错误和在数据判断与解读中出现的正常差异。2002 年，美国国家科学基金会又在此基础上补充三个内容，确定学术不端行为必须要有以下情况：①必须明显偏离相关学术界公认的行为准则；②学术与研究不端行为是行为人蓄意知情或鲁莽造成的；③必须有充分的证据证明学术不端行为。以上定义都明确禁止“捏造、篡改和剽窃”，这通常被称为 FFP（fabrication falsification plagiarism）核心因素，并已成为许多部门和机构定义学术不端行为的共同特点。此后，一些学术团体、大学和研究机构通过直接引用美国公共卫生署和国家科学基金会的定义，或将它们作为修改的蓝本，分别拟定了各自对学术不端行为的含义。

在我国，科技部 2006 年颁布的《国家科技计划实施中科研不端行为处理办法（试行）》，对科研不端行为的定义是“违反科学共同体公认的科研行为准则的行为”。2007 年

1 月 16 日中国科协七届三次常委会议审议通过的《科技工作者科学道德规范(试行)》第三章对学术不端行为下了明确的定义:“学术不端行为是指在科学研究和学术活动中的各种造假、抄袭、剽窃和其他违背科学共同体惯例的行为。”2007 年 2 月 26 日中国科学院发布的《中国科学院关于加强科研行为规范建设的意见》将科研不端行为概括为六个方面:①在研究和学术领域内有意做出虚假的陈述;②损害他人著作权;③ 违反职业道德利用他人重要的学术认识、假设、学说或者研究计划;④研究成果发表或出版中的科学不端行为;⑤故意干扰或妨碍他人的研究活动;⑥在科研活动过程中违背社会道德。2009 年,教育部又针对高校学术不端行为频增的事实,专门下发了《关于严肃处理高等学校学术不端行为的通知》,指出高等学校对七种学术不端行为必须严肃处理:①抄袭、剽窃、侵吞他人学术成果;②篡改他人学术成果;③伪造或者篡改数据、文献,捏造事实;④伪造注释;⑤未参加创作,在他人学术成果上署名;⑥未经他人许可,不当使用他人署名;⑦其他学术不端行为。许多高校也以此为鉴,分别制定了适合本学校要求的治理学术不端行为的具体条款并行之有效地付诸实践。可见我国对学术不端行为的治理,逐步走向明确化、规范化、合理化。

2.5.2　大学生学术不端行为的表现

目前在大学生的本科和研究生学习过程中尚有六种不端行为:①抄袭、剽窃网络资源中的已发表论文,并不加以标注;②利用网络现有文献资料编造、篡改数据资源,为己所用;③肆意盗用他人的学术观点,不标明出处;④进行论文买卖交易;⑤利用中介机构,进行论文代写代发;⑥利用手机等其他电子资源进行考试作弊;⑦课程作业、实验报告、生成实习报告、课程设计、大学生创新项目申报书撰写、普通论文撰写与发表、毕业论文撰写与答辩等活动中故意抄袭与复制、恶意分享与剽窃较严重。这七种学术不端现象在大学生的学习过程与学术生涯中呈渐进式滋长与蔓延。此外大学生对于他人的学术不端行为也表现出事不关己,听之任之的消极态度。根据《当代大学生利用网络学术资源不端行为的调查》课题组调查结果显示,在 1492 份有效调查问卷中仅有 22.8%的学生对其他同学的学术不端行为表现坚决抵制,有 36.1%的大学生表示可以容忍和接受,32.8%的大学生表示无所谓,甚至有 8.3%的大学生表示支持。这种漠视、纵容的态度折射出大学生学术诚信缺失及学术素养低下的现实状况。在对浙江省内五所高校的在校大学生进行学术诚信问题的调查中,有 62%的大学生认为考试作弊是非常普遍的,有 83.9%的大学生认为作业抄袭是普遍存在的现象,70%的大学生对其他同学考试作弊行为视而不见,更有甚者表示有需要时自己也会作弊,这一比例高达 60.4%。这些调查结

果充分说明急功近利、心浮气躁、缺乏诚信是现时期大学生学术行为的主要特征。

当前大学生存在的不端行为有三个方面的影响因素。第一,个人利益的驱使是引起大学生学术不端行为泛起的内在诱因。伴随市场经济的发展,一种片面追求物质财富的社会氛围日益形成,熏染着大学生的社会价值观向多元化、世俗化、功利化转变,在物欲横流的现实生活中,权钱交易、权权交易、权色交易在国家各个层面反腐倡廉与惩戒腐败的高压态势下仍然屡禁不止,并逐步涉足校园净土,权学交易随之而来,这种歪风邪气的蔓延,使部分大学生放弃了对"发愤图强"、"为中华崛起而读书"的崇高梦想与学术精神的追求,忽视了学习知识与创新创业的社会责任,过分注重结果而轻视学业,为获取学位不择手段,把学术当成获取个人私利的工具。第二,缺乏应有的信息检索素养教育、诚信教育与学术素质塑造。在浙江省内五所高校大学生学术诚信问题的调查问卷结果中,有66.9%的学生认为诚信缺失的原因是学术道德教育薄弱,有39%的大学生认为学校没有开设过学术规范课程,34.2%的学生不清楚学校是否开设过学术规范课程。类似这五所高校的调查结果在全国各高校内普遍存在,甚至有过之而无不及。可见高校对大学生学术道德教育的重视程度不够,大学生对学术规范缺乏了解和认知,以至于不端行为愈演愈烈。第三,大学生学术能力水平低。许多大学生都知道抄袭、剽窃、作弊等行为是学术道德失范的典型表现,但仍侥幸尝试,这种明知故犯的背后,是由于学术水平低下,学术知识匮乏,学术创新能力薄弱,但又想得到良好的学术成果和考试成绩,只好不择手段去"复制"或"窃取"他人成果。现在大学生的普遍弊病是注重吃喝享乐,缺乏刻苦钻研与独立思考的精神,在"填鸭式"教学模式及浮躁的社会风气影响下,形成了懈怠、懒惰的学习态度和骄纵、奢靡的生活作风。这种校园风气使大学生对学术不端行为产生麻木而纵容,作弊、抄袭等不端行为屡禁不止。

2.5.3 信息检索素养教育对大学生学术不端行为的作用

(1)利用信息检索原理及其技术应用,反制大学生学术不端行为。以往大学生主要是通过在图书馆借阅书籍和查阅文献获取相关信息和知识,而网络技术及其信息资源建设快速发展却改变了这一传统的方式。虽然网络的普及给人们带来了便捷、快速获取信息的手段,但是通过网络手段唾手可得的丰富信息,却成为了抄袭与剽窃的主要来源。丰富的资料以及简单的"复制"与"粘贴"就能使人们不费吹灰之力"拼凑"成一篇论文。特别是现今对外文资料的获取较之从前更为容易,这就为一些人在写论文时提供了一种新的抄袭手段,即将外文资料翻译成中文不加注释地引用到自己的论文中。

高校需要用信息检索的原理,独立建设或引进"学术不端检测信息系统",利用信息检

索基本原理构建庞大的“学术不端检测网络系统”，对学生的课程大作业、实验报告、课程设计报告、生成实习报告、大学生创新创业项目申报书、毕业论文等，通过抽查或全面复查方式在“学术不端检测网络系统”中检测。检测的基本原理包括文本字符信息检测、图像图表信息检测、数学原理与化学分子式检测、音频信息检索、视频信息检测、跨库跨平台检测等技术实现学生学习过程档案的无缝检测与对接。通过信息检索原理及其技术应用的力量对学生的学术不端行为进行全覆盖、无死角治理，反制大学生学术不端行为。

(2) 加强建设对大学生的学术不端行为审查与惩处力度。高校应加强建设有关大学生的学术不端行为审查制度，对学术行为进行严格把关。虽然 2004 年教育部社会科学委员会制定的《高等学校哲学社会科学研究学术规范(试行)》明确规定“不得以任何方式抄袭、剽窃或侵吞他人学术成果”。各个高校则应该积极制定实施“学术规范”的制度性办法，建立起完善的审查制度，一旦发现存在抄袭、剽窃等学术不端问题，应立即加以惩处。首先，高校可以将“学术规范”与“学位证”挂钩，对有抄袭剽窃他人成果的学生予以惩罚，才能实现教育部制定“学术规范”政策的真正目的，同时对大学生的学术不端行为形成零容忍制度，包括事前发现、事中发现和毕业后发现，均属惩处的制度范畴之列，从制度上根本保障大学生的学术行为回归到良性轨道上来。

(3) 通过信息检索素养教育，提高学生的科研能力。信息检索素养教育的重要目的就是“站在巨人肩膀上”，通过全面的信息检索，把握研究项目的立项依据、研究可行性、研究方法、技术路线、研究计划与研究预期的科学性，少走弯路，达到事半功倍的研究目的。通过信息检索素养教育，结合学生的专业学习，提高学生的科研能力，这是对大学生的学术不端行为进行根本治理的治本之策与必然出路。对大学生的学术不端行为审查与惩处规范无论有多么完善，都需要“通过信息检索素养教育，提高学生的科研能力”作为坚实基础。信息检索素养教育能够逐步建立并增强大学生自主学习、发现与探究学习、协作与研究性学习的过程与能力，从根本上建立学生摒弃各种学术不端行为的强烈自信心与能力基础。

本 章 小 结

日益发达的信息技术手段和无处不在的网络化环境，已经可以让人们轻而易举而又不露痕迹地检索、获取、共享和利用各种信息资源或信息产品，这就为各种信息化犯罪创造了条件，人们的信息检索与利用活动迫切需要形成更加广泛和深入的知识产权法律意识，也是信息检索道德的内在要求。

由于信息是知识产权活动的一种客观反映形式，而当代大学生作为信息社会信息检索与利用过程中最具活力的生力军，需要具备较高的知识产权法律意识，尊重知识产权，杜绝、避免知识产权侵害和各种网络化信息犯罪的发生，共建公平、开放、和谐与守法的信息化与网络化环境。这不仅是顺利和合法开展信息检索与利用活动的前提与根本保证，也是当代大学生信息检索素养教育的内在要求。

信息社会的三个核心要素：信息技术带动高新技术发展，信息产业促进传统产业结构加速调整，信息资源引导经济集约化。

信息是事物运动的一种状态与方式，是物质的一种属性。信息有无穷性、可辨识性、可转换性、可存储性、可传递性、可分享性等重要特征。

知识产权主要是指人们对其从事智力活动而产生的成果所依法享有的专有权利，是一种无形财产权。知识产权是人类的发明创造、智力活动成果和法律活动的结合与交叉，是人们依据国家法律对自己的智力活动而获得的成果所享有的权利。对于大学生常常检索与利用的专著、学位论文、专利、标准、学术研究论文等信息作品以及网络原创性信息都属于知识产权范畴，因为信息是知识产权活动的一种反映，也是知识产权现象的表述，体现了知识产权的主体内容。

作为当代大学生，通过各种信息检索技术手段以满足自身不断增长的信息需求，是时代的要求与必然趋势，但前提是不能侵害他人或组织的正当知识产权利益，不得干扰或危害和谐的信息资源利用环境和信息共享秩序，不得侵害个人隐私信息。在获取与利用信息的同时，信息检索与利用者的检索手段和利用方法也必须置于法律和道德允许的范围内。

信息社会中，信息的海量增长、网络化关联、易检索性、易复制性、易扩散性等特点，决定了完全依靠刚性约束为主的法律手段来规范信息检索、获取与利用者的行为是不现实的，也是比较困难的。遵守信息道德，是信息化时代和网络化环境下全部个体与组织的共同利益需求，信息道德失范行为不符合他人的利益和公共利益，归根到底也有损于自身利益，所以需要信息检索道德的培养和自律的逐渐形成。

知识产权对信息资源的检索与共享是有制约作用的，同时又起着促进作用。作为大学生，信息检索的主要目的是为了自主学习、发现与探究学习、协作与研究性学习、课题与项目研究活动及其生活与休闲娱乐等活动服务，在满足信息需求的基础上，提高学习、研究、工作和生活的效率与质量。但是，大学生侵犯他人知识产权、侵害别人智力劳动成果与学术不端行为却屡屡发生，有着很多形成的环境因素。

信息检索素养教育对大学生学术不端行为起着主要作用。通过利用信息检索原理及

其技术应用,反制大学生学术不端行为。通过加强对大学生的学术不端行为的审查与惩处力度,从制度上根本保障大学生的学术行为回归到良性轨道上来。通过信息检索素养教育,逐步建立并增强大学生自主学习、发现与探究学习、协作与研究性学习的过程与能力,从根本上建立学生摒弃各种学术不端行为的自信心与能力基础。

本章思考与练习题

1. 什么是信息社会? 它的主要特点是什么?
2. 信息社会的三个核心要素是什么?
3. 信息的含义与基本类型有哪些?
4. 信息有哪些主要特征? 分别举例说明。
5. 知识产权的内涵? 其主要特点包括哪些方面?
6. 知识产权的性质或本质是什么? 请举例说明。
7. 知识产权有哪些主要内容?
8. 信息的财产权属性如何体现?
9. 什么是信息公开? 知情权含义是什么?
10. 信息检索与利用过程中,如何识别国家秘密、商业秘密和个人隐私等信息?
11. 当代大学生如何在信息检索与利用过程中逐渐培养道德自律?
12. 举例说明信息检索与利用同知识产权保护的相互影响作用。
13. 哪些行为属于大学生学术不端行为? 在信息检索与利用过程中你认为该如何避免?
14. 大学生学术不端行为的概念如何界定? 有哪些表现与成因?
15. 信息检索素养教育对大学生学术不端行为有何作用?

第3章　信息检索的基本知识

3.1　信息检索的含义

3.1.1　检索的概念

信息检索起源于对文本信息和印刷资料的情报检索，开始于20世纪50年代初期。1954年，美国海军军械试验站图书馆利用IBM-701电子计算机建立了世界上第一个信息检索系统，用于情报服务。1959年，H. P. 卢恩（Luhn）利用IBM-650计算机对文献信息进行统计分析，实现定题情报检索服务。20世纪60年代，在图书情报工作中广泛利用计算机脱机批处理系统进行情报检索。1962年，美国M. M. 凯瑟尔进行了世界上最早的联机信息检索试验。1964年，美国系统发展公司（SDC）成功研制"书目信息实时共享在线检索"（on-line retrieval of biographic information-time shared，ORBIT）软件。20世纪70年代以来，人们对信息检索进行了大量的理论和应用研究。联机信息检索系统除了上述的Orbit之外，还有美国国家医学图书馆的Medline系统、美国洛克希德公司的Dialog系统。与此同时，法国、英国、日本、加拿大也先后建立了联机信息检索系统，如欧洲空间组织情报检索中心的ESA-IRS系统。

20世纪90年代以来，以互联网技术发展为支撑的网络信息资源迅猛增长，人们将发展较为成熟的专业性的文本信息检索原理与方法移植到Internet网上，这大大促进了信息检索的发展、推广与普及化，使得信息检索从相对封闭、稳定一致、由独立数据库集中管理的信息内容扩展到开放共享、动态更新、传播快速、松散管理的Web信息世界。

信息检索是一个外来词汇，源于英文的"Retrieval"，其英文近义词是"Search"和"Query"，翻译成中文是"查找"或"查询"的意思。检索是指从图书文献、学术期刊、专题数据库、网络信息系统、学科网站等各种信息资源集合中，利用一定的方法与技术查找符合自己需要的信息或资料，从而满足自身信息需求的过程。

广义的检索是指将信息按一定的方式组织和存储起来，并根据用户的信息需要查询出有关信息的活动与过程，所以它的全称又叫信息存储与检索。检索概念的广义内容包

括信息存储与信息检索的集成化过程。例如信息集合中某一信息的存储规范与信息用户的检索规则与需求表达一致，信息集合就能成功提取该信息给用户，否则信息检索与获取过程就会失败。狭义的信息检索仅指通过该过程的后半部分，即从信息集合中查找并获取所需信息的过程，相当于人们所说的信息查询过程或查询活动。

3.1.2　信息检索的含义

信息检索(information retrieval)术语最早产生于美国学者 Calvin Mooer 在 1948 年的 MIT 硕士论文。Information Retrieval(IR)：简单地说是从文档集合中返回满足用户需求的相关信息的过程。作为一门学科领域，是研究信息的获取(acquisition)、表示(representation)、存储(storage)、组织(organization)和访问(access)的一门学问。

用户需求(user need，UN)：指的是用户需要获得的信息。严格地说，UN 只存在于用户的内心。UN 提交给检索系统时称为查询(query)，查询通常用文本来表示，对同一个 UN，不同人不同时候可以构造出不同的 Query 表达式。

文档(document)：检索的对象。可以是文本，也可以是图像、视频、语音等多媒体文档。相应称为文本检索(text retrieval)、图像检索(image retrieval)、视频检索(video retrieval)、语音检索(speech retrieval)、多媒体检索(multimedia retrieval)。文档可以是无结构的、半结构的、有结构的。

文档集合(collection)：所有待检索的文档构成的集合，也称为知识库(repository)、语料库(corpus)或数据库(database)。

信息相关和信息相关度(relevant、relevance)：相关性概念是信息检索的核心。信息检索的主要目标就是检索出所有与用户查询相关的文档。相关取决于用户的知识积累与信息需求判断，是一个主观的概念。不同用户做出的判断很难保证一致；即使是同一用户在不同时期、不同环境下做出的判断也不尽相同。“相关性”的研究，从 20 世纪 30 年代至今已经有 80 多年的历史，期间两个主要的流派分别是面向信息系统的相关性研究以及面向信息用户的相关性研究。研究的高峰分别集中于 20 世纪 60 年代至 70 年代前期，以及 80 年代中后期至今的两个阶段。相关性是动态的、多维的、认知的以及可测度的等观点已经成为学术界的共识。

概括地说，信息检索就是从非结构化的信息集合中发现、查询并评价与用户需求相关的信息。相应地，信息检索系统就是用来实现信息检索功能的计算机软件系统或网络信息系统。

这里要强调的是，与数据库系统处理的结构化信息不同，信息检索处理的是“非结构

化信息”。

什么是“非结构化信息”呢？一篇新闻就是一条非结构化信息，新闻中会出现一些人名、地名、机构名等实体，以及这些实体之间的关系（比如某人是某地某机关的负责人），还有与这些实体相关的事件（比如某人访问了某地）。但这些人、事、物、关系和时间并不像关系数据库的二维表中存放的信息那样，被精确地分割并严格地存放在合适的字段或记录中。这种在现实世界中自然存在的模糊而带有歧义且没经过规格化的信息被称为“非结构化的”信息。

现实世界中存在着大量的非结构化信息，除文本外，还有图像、图形、语音、视频等多媒体信息。文本中又有各种各样的类型，如网页、邮件、博客、论坛上的帖子、聊天记录、短信、论文、报告、技术标准、法律文档、统计报表等，不同类型的文本各有不同的特点，比如论坛上的帖子往往非常口语化，存在大量的别称、省略语等现象，给信息检索带来很大的挑战。

要处理好非结构化文本，就要尽可能地从非结构化信息中找出一些结构来。所谓的“非结构化信息”并不是真的没有结构，只是其结构不是显性存在的，而是隐含的，要找出其中的结构需要运用由浅到深的各类文本检索处理技术。比如，中文分词技术就可以把词语从句子中分割出来，而隐性语义分析技术则可以从词汇与文档关系的信息挖掘中发现文本的深层结构。

用户的信息检索过程可以描述为：用户提交信息需求的查询条件，信息检索系统根据该查询条件在文档集中检索出与其相关的文档子集，对这些相关文档子集中的文档按照与查询条件的相关性度进行排序，最后返回给用户有序的文档子集。信息检索的形式化描述如下。

定义：假设信息检索模型是一个四元组$\{D,Q,F,R(d_j,q)\}$。其中D是文献集中的一组文献逻辑表示，称为文献表示；Q是一组用户信息需求的逻辑视图（表示），这种视图（表示）称之为查询；F是一种机制，用于构建信息表示、查询及它们之间关系的模型；$R(d_j,q)$是排序函数，该函数输出一个与查询Q和信息表示D有关的实数，这样就在信息文档之间根据查询R，定义了一个顺序。

3.1.3 信息检索用户的基础素养

对于信息检索的用户而言，信息检索通常要具备四个基础素养。

（1）用户信息意识。用户信息意识是信息检索的前提。所谓信息意识，简单地说，是人们利用信息系统获取所需信息的内在动因，具体表现为对信息的敏感性、选择能力和消

化吸收能力。信息意识含有信息认知、信息情感和信息行为倾向三个层面。信息素养(素质)(information literacy)一词最早是由美国信息产业协会主席 Paul 在 1974 年给美国政府的报告中提出来的。他认为:信息素质是人们在工作中获取信息、学习信息技术、利用信息资源解决问题的能力。

(2) 信息源掌握。信息检索的基础是信息源(信息的来源)掌握。信息源的构成:按文献载体分为印刷型、缩微型、机读型、声像型和网络型信息源;按文献内容和加工程度分为一次信息源、二次信息源与三次信息源;按出版形式分为图书、报刊、研究报告、会议信息、专利信息、统计数据、政府出版物、档案、学位论文与标准信息等信息源(它们被认为是十大传统信息源,其中后八种被称为特种文献)。对于学习者而言,学习与研究型信息源主要分布在教育类图书、专业研究期刊、学位论文等不同类型的出版物及其数据库中。

(3) 信息获取能力。信息获取能力是信息检索的核心元素。获取能力要求:了解各种信息来源;掌握检索语言、熟练使用检索工具、能对检索效果进行判断和评价。判断检索效果的两个指标,查全率=被检出相关信息量/相关信息总量(%);查准率=被检出相关信息量/被检出信息总量(%)。

(4) 信息共享与利用。信息共享与利用是信息检索的关键所在。社会进步的过程就是一个信息不断的生产→存储→传播→再生产的过程。为了全面、有效地利用现有信息资源促进我们的学习、工作和生活效率与质量,各行各业信息检索的需求量与检索活动的比例与日俱增。

3.1.4　信息检索的领域与范畴

信息检索作为一个学科或研究领域,是信息学领域的一个重要分支。

信息检索的基本知识与原理来源于计算机科学、数学、信息科学、语言学、信息论、图书馆学、情报学、认知心理学、语言学、统计学、管理学等学科,现在已经扩充拓展到了财经、化学、物理学、航空航天等领域,又随着人工智能、认知科学、计算机技术、互联网技术、大数据挖掘、神经科学、多媒体技术、云计算、智慧城市、智慧社区和智慧教育等新兴领域的不断延伸与交叉融合,当今信息检索将逐渐适应人脑的思维方式,实现智能、高效、快速而灵活的信息检索与共享,最终达到随心所欲查找、快速获取和高效利用信息的目的。

信息检索的研究日益与数学、计算机科学、系统学、语言学、信息论等学科紧密结合起来,大大扩展了自身研究领域和研究队伍,数学、通信、计算机科学、管理学、语言学等领域的许多学者与专家也加入到信息检索研究领域,形成一种学科深度融合的新局面。在计算机技术领域的新硬件、新软件、新技术、新方法的支持下,信息检索的研究水平也已从现

象描述阶段进入大规模试验阶段,新的文摘方法、索引语言、索引方法、智能分析技术、高层次检索系统应用产品及其检索质量评价方法等分支研究领域不断涌现,从而为建立起信息检索技术理论和研究方法体系奠定了基础,为满足广大信息用户在信息化社会环境中日益增长的信息需要奠定了基础。

3.1.5 信息检索的类型

随着信息检索逐步被人们认识、掌握、利用以及人们对信息不断增长的需求,信息检索类型也得到不断丰富。按照不同的标准,信息检索可划分为不同类型。按照信息检索对象以及信息表现形式的不同,可分为简单的纸质载体信息形式和较复杂的电子媒体形式。按照信息检索内容可分为书目检索、数据检索、事实检索、全文检索、图像检索和音频检索。

(1) 书目检索(bibliography retrieval),它是以纸质载体为检索对象的信息检索。即检索内容存储于书目、索引、文摘等纸质文献(例如图书馆馆藏书目、美国的科学文摘等)中,它是原始文献信息(图书、期刊、报纸等)的外表特征与内部特征的简化描述,是传统文献资源的"替代物",信息用户通过检索获得的是与检索课题有关的一系列文献线索,然后通过查阅和阅读决定取舍。书目检索相对于全文检索、数据检索、事实检索而言是产生较早的检索形式,其发展也较快,比如各个图书馆的"馆藏书目检索数据库"。

(2) 数据检索(data retrieval)。数据检索具有数量性质,它是以数值形式表示检索内容的信息检索形式,即其中存储大量数据以便查出专门的数据资料,这种专门的数据经过专门的测试、评价及筛选,用户检索到的各种数据可直接使用或进行定量分析,例如各种统计行业、金融业、证券业等行业的数据检索与分析库。

(3) 事实检索(fact retrieval)。事实检索是以信息资源中抽取的事实为检索内容的信息检索,它从检索系统存储的各种原始信息资料中查找特定的事实材料为检索目的。事实材料指出事物的性质、定义、原理与发生的地点、时间以及因果关系等,例如各类报业、电视等媒介集团的新闻门户网站提供事实类信息检索的资源十分丰富。

(4) 全文检索(full-text retrieval)。全文检索是原始信息所含的全部信息,即以整篇文章或整体图书为检索内容的检索需要,检索的内容可以是全文,也可以是部分内容,并可以进行各种频率的统计和内容分析,它通常用自然语言表达检索需求。全文检索是现代检索的发展方向,它与书目检索最根本的区别是它对最终需求的信息进行了最全面的描述。例如传统的科学文摘或搜索引擎,只是提供了最终需求信息的"线索"(例如简介或链接地址),而全文检索直接提供原始全文信息。

(5) 图像检索(graphic retrieval)。图像检索是指以有关人物事物的图片、图像和图文信息为检索内容的检索活动。它利用计算机数据库存储图像信息,以图形的色彩、纹理、轮廓等特性为检索方法和获取依据的系统。这类检索系统在教学、科学研究、医疗诊断、旅游参观以及各种宣传广告领域发挥着重要作用。视频信息检索也属于图像检索的大范畴,因为连续的视频流首先要分割为图像后才能进行特征化检索处理。

(6) 音频检索。音频检索是以波形声音为对象的检索,这里的音频可以是汽车发动机声、雨声、鸟叫声,也可以是语音和音乐等,这些音频都统一用声学特征来检索。使信息用户能从大型音频数据库中或一段长录音中找到感兴趣的音频内容是音频检索的目的。音频数据的训练、分类和分割方便了音频数据库的浏览和查找,基于听觉特征的检索为用户提供高级的音频查询接口。音频检索就是针对广泛的声音数据的检索,需要检索的音频可以包含语音和音乐,但是采用的是更一般性的声学特性分析方法(包括音频特征分割、提取与统计等处理)。

3.2　信息检索涉及的相关支撑领域

信息检索是一门多学科交叉的应用领域。信息检索的对象包括文字、图像、图表、音频、视频等多种媒体信息,信息检索需要利用各类媒体处理技术(比如自然语言处理、图像处理、语音处理、视频处理等)对信息进行加工,找出一定的结构,为信息检索与获取提供支持。信息检索通常要面对海量数据,普通台式机的处理能力远远不够,并行与分布式处理、云计算、大数据处理等新的理论与方法在这个领域大有用武之地。数据库和数据挖掘被用来解决结构化信息检索与知识发现问题,它们已取得的成果对文本信息检索与文本信息挖掘都有直接的借鉴作用。知识管理、情报学、社会学等偏重人文与管理的学科从不同的角度使用信息检索技术并从中获益。

(1) 自然语言处理。自然语言处理是利用计算机技术处理语言信息,其目的是让计算机能够“理解”人类的语言——自然语言。对于信息检索来说,仅仅停留在处理表层文本信息是远远不够的,字符层面的匹配与相似度计算并不能帮助计算机理解检索对象文本的“含义”,也不能深入理解用户的检索意图,检索出的结果很有可能偏离用户的实际信息需求。要提高检索系统自身的智能化水平,以及检索系统人机交互界面的自然度,就需要不断地将自然语言处理结合到文本信息检索中来。

(2) 分布式计算。Internet 构成了人类历史上最大的开放性信息平台,拥有海量的数据。面对巨大的文本数据、大量的检索请求和用户对检索时间的严格要求,信息检索的效

率与质量必然成为一个亟待解决的问题，依靠单台计算机不可能完成这样的任务，必须依靠分布式信息检索技术才能解决。事实上，几乎所有实用的大型搜索系统都采用了分布式的体系结构来解决信息检索中的效率问题。

（3）数据库技术。数据库和信息检索俨然一对姐妹。与信息检索不同，数据库的处理对象是结构化信息。数据库技术已经有比较完整的理论基础，而信息检索的经验性比较强，理论基础相对薄弱，需要进一步借鉴数据库中的一些成熟理论。信息检索中的信息抽取技术旨在把非结构化数据转化为结构化数据，以数据库形式存储和处理，因而，信息检索与获取问题就可以转化为数据库查询问题。

（4）数据挖掘。数据挖掘一般是针对数据库进行的，借鉴到信息检索过程中就成为文本挖掘。面向非结构化数据的文本挖掘，帮助用户对互联网上庞杂的信息进行综合分析，找出这些信息背后所蕴含的规律和用户倾向性，找出信息的本质含义，提升搜索质量。其中对用户的信息检索日志进行数据挖掘能够从总体上观察分析用户的检索行为和需求倾向，也能够针对每个个体用户的需求提供个性化服务。

（5）情报学。情报学是研究情报的产生、传递和利用规律的学科，是研究情报流通过程和情报系统保持最佳效能的一门学问。它能帮助人们充分利用信息技术手段，提高情报产生、加工、存储、流通和利用的效率。信息检索和情报学有紧密的历史渊源，情报学的理论对信息检索系统的设计有指导作用。

（6）社会学。社会学是研究社会发展现象和规律的科学。随着搜索引擎技术的使用越来越广泛，社会学家通过对众多用户使用搜索引擎的行为（比如浏览了哪些网页、输入了哪些查询词、网页点击量与停留时间、输入检索关键词的数量与频率等用户行为特征）进行分析和统计，来研究社会心理、行为或群体信息交流的状态和趋势，为信息资源聚类、用户特征聚类和提升检索质量起着重要作用。

（7）云计算。云计算（cloud computing）是分布式计算（distributed computing）、并行计算（parallel computing）、效用计算（utility computing）、网络存储（network storage）、虚拟化（virtualization）、负载均衡（load balance）、热备份冗余（high available）等传统计算机和网络技术发展融合的产物。美国国家标准与技术研究院（NIST）定义：云计算是一种按使用量付费的模式，这种模式提供可用的、便捷的、按需的网络访问。这种资源池称为“云”。“云”是一些可以自我维护和管理的虚拟计算资源，通常是一些大型服务器集群，包括计算服务器、存储服务器和宽带资源等。云计算将计算资源集中起来，并通过专门软件实现自动管理，无须人为参与。用户可以动态申请部分资源，支持各种应用程序的运转，无须为烦琐的细节而烦恼，能够更加专注于自己的业务，有利于提高效率、降低成本和技

术创新。

(8) 大数据。大数据(big data)指无法在可承受的时间范围内用常规软件工具进行捕捉、管理和处理的海量信息和数据集合，需要新处理模式来处理日益膨胀的海量信息和数据集合，这种模式需要具备更强的决策力、洞察发现力和流程优化能力，来生成高增长率和多样化的信息检索、共享与利用资产。在维克托·迈尔-舍恩伯格及肯尼斯·库克耶编写的《大数据时代》中：大数据指不用随机分析法(传统的信息抽样调查)这样的捷径，而采用所有的数据进行分析处理。大数据的5V特点(IBM提出)：Volume(大量)、Velocity(高速)、Variety(多样)、Value(价值)和Veracity(真实性)。

我国高度重视大数据发展。经李克强总理签批，2015年9月国务院印发《促进大数据发展行动纲要》，系统部署大数据发展工作。明确推动大数据发展和应用，在未来5～10年打造精准治理、多方协作的社会治理新模式，建立运行平稳、安全高效的经济运行新机制，构建以人为本、惠及全民的民生服务新体系，开启大众创业、万众创新的创新驱动新格局，培育高端智能、新兴繁荣的产业发展新生态。一要加快政府数据开放共享，推动资源整合，提升治理能力。大力推动政府部门数据共享，稳步推动公共数据资源开放，统筹规划大数据基础设施建设，支持宏观调控科学化，推动政府治理精准化，推进商事服务便捷化，促进安全保障高效化，加快民生服务普惠化。二要推动产业创新发展，培育新兴业态，助力经济转型。发展大数据在工业、新兴产业、农业农村等行业领域应用，推动大数据发展与科研创新有机结合，推进基础研究和核心技术攻关，形成大数据产品体系，完善大数据产业链。三要强化安全保障，提高管理水平，促进健康发展。健全大数据安全保障体系，强化安全支撑。

3.3　信息检索的前沿与热点问题

3.3.1　信息检索的发展趋势

如何快速、准确、全面地检索并获取到所需信息，在信息时代无论对于个人或组织都十分重要。近年来，信息检索取得了飞速的发展，特别值得一提的是，中文全文检索技术的发展非常迅速，并且国内自主开发的产品取得了绝大部分的市场份额，这对于一个以核心技术为竞争优势的领域是非常难能可贵的。著名的全文检索系统TRS在政府、企业、媒体和教育领域都取得了卓越的成绩，市场占有率在70%以上。目前全文检索的技术已经比较成熟，正在得到广泛应用，而多媒体检索和智能检索的研究与应用还有一定距离。主要面临的问题是音频与视频检索技术及其人工智能检索技术的发展还不尽如人意。在

网络搜索引擎方面主要是向集成化、专业化方向发展，单一搜索引擎过渡到集成化的多元搜索引擎和专业化的检索与信息服务领域。随着网络信息量的迅猛膨胀，对智能化检索工具的智能化程度提出了更高的要求。智能化程度高的检索工具在竞争中将明显地处于有利的地位。总的来说，信息检索技术有以下一些发展趋势。

1. 可视化和多样化

信息检索中的可视化，是将数据库中不可见的语义关系用图像形式可视化显示，并可视化表达用户检索过程。而网络信息检索多样化首先表现在可以检索的信息形态有文本、声音、图像、动画等信息资源。目前网络信息检索的主体是文本信息，基于内容的检索技术和语音识别技术的发展，将使多媒体信息的检索变得逐渐普遍。基于内容的检索是指根据媒体和媒体对象的内容及上下文联系在大规模多媒体数据库中进行检索。它的发展目标是提供在没有人参与的情况下能自动识别或理解图像重要特征。目前，基于内容的多媒体信息检索的主要工作集中在识别和描述图像的颜色、纹理、形状、空间关系上，对于视频数据，还有视频分割、关键帧提取、场景变换感知以及故事情节重构等问题。由此可见，这是一门涉及面很广的交叉学科，需要利用图像处理、模式识别、计算机视觉、图像理解等领域的知识作为基础，还需从认知科学、人工智能、数据库管理系统、人机交互、信息检索处理技术等领域引入新的媒体数据表示和数据模型，才可能设计出可靠、有效的检索算法、系统结构以及友好的人机交互界面。多样化的第二个表现是检索工具向多国化、多语种化方向发展。多样化的第三个表现是网上检索工具的服务形式多样化。

2. 集成化

目前种类繁多的网络数据库都缺乏统一的数据描述标准，但是信息查询方式大相径庭。数据类型不同，信息系统返回给用户的检索结果也不尽相同。因此，用户在使用各种信息资源库之前，必须花费大量时间对其检索方法、检索系统逐一学习和掌握，同时用户在对不同的数据库进行检索时还必须切换不同的检索交互界面，采用不同的专用阅读器进行不同格式的数据转换和阅读。鉴于这些诸多的不便之处，集成化已成为信息检索服务的一大发展趋势。例如数字图书馆通过提供集成检索机制，方便用户从一个检索界面同时检索数字图书馆的所有资源，避免重复多次登录、多次检索的麻烦。用户利用集成信息检索服务时，所面对的是“一步到位”式的一站式计算机检索界面，而后台则是整体化的信息资源保障体系。目前国内的 CNKI、重庆维普等数据库都正在尝试实现本公司开发的不同数据库之间的集成检索；而国外如 OCLC 已开发出帮助数字图书馆建立集成信息检索的工具——OCLC Web Express，使用它可以把 OCLC 提供的各种服务与数字图书馆的其他电子资源集成在一个界面上，便利用户对信息资源的集成检索；可把馆内外、远

程和本地的信息资源统一到独立的集成服务界面，把 OCLC 和非 OCLC 的信息资源集合到统一的搜索界面上。

3. 个性化

为了提高用户满意程度，将用户所需信息准确返回，信息检索必须向个性化方向发展。信息检索个性化的核心是跟踪分析用户的检索行为和个性化信息检索需求，充分利用这些信息来提高用户的检索效率。通过检索行为分析提高检索效率的途径有两种："群体行为分析"（比如一些数据库中列出的"热门关键词"就是这种分析的运用结果）和"个性化检索"（通过积累用户的检索个性化数据，使用户的检索更深入、更精确）。信息检索服务个性化还表现在实现用户检索习惯的个性化定制。在用户检索网络资源或数据库信息过程中，往往由于拥有的检索知识和所处领域不同，其检索操作和检索习惯也有所差异。例如，初学者习惯于简单检索，而专业人员则习惯于使用高级检索。此外，不同用户对检索结果的选取原则和排序方法也不尽相同。例如，有的人希望按相关度排序，有的则偏向于按网站的点击量，如此种种都反映了用户的个性化需求。因此，个性化检索服务还应包含对习惯性检索机制的定制，主要应包括检索工具定制，即选用常用的搜索引擎和数据库；检索表示方式定制，可选择常用的检索式（如布尔逻辑检索式中的"与"、"或"、"非"等逻辑查询）为默认方式；检索结果处理定制，可对检索结果的相关度计算标准、输出格式、排序方式等进行定制。

4. 智能化

准确的信息检索工具应建立在对收集信息和检索请求的理解之上，也就是说必须处理语义信息。传统的信息检索是被动式的，而利用智能代理技术进行主动信息检索则逐渐成为这一领域的焦点。其中通过对用户的信息需求规划、检索意图、需求兴趣或专业方向进行推理，预测并为用户提供有效的检索反馈。信息检索智能化使用自动获得的知识进行信息收集过滤，并自动将用户感兴趣的信息通过电子邮件、社交网络平台或其他方式提交给信息用户。智能检索工具由于将信息检索从目前基于关键词层面提高到基于知识（或概念）本体层面，对知识有一定的理解与处理能力，能实现智能分词与切词技术、同义词与近义词辨析、概念检索、短语识别以及机器翻译等，从而使信息检索更具有智能化和人性化特征。

5. 基于网格的信息检索

网格的构想来源于电网，就像人们使用电器设备，能量从电网中迅速传输到所需设备中。网格实际上就是利用互联网将分散于不同地域的计算机组织起来，成为一个虚拟的超级计算机，每台参与的计算机就是一个"节点"，成千上万的节点纵横交错，构成一张"网

格”或“计算机云”。它可以连接和统一各类不同远程资源，实现互联网上所有资源的全面连通与透明化，在动态的、异构的虚拟组织间实现网络虚拟环境上的资源共享和协同工作，从而消除信息检索与共享过程中的信息孤岛和资源孤岛。美国科学家、网格运算项目领导人之一的 Lan Foster 曾描述：网格是构筑在 Internet 上的一组新兴技术，将高速互联网、电脑、大型资料库、传感器、远端设备等融为一体，为科技人员与普通网民提供更多精准资源、增值功能及个性服务。由于网格的特性与信息检索存在共同点，利用网格中的一些原理可以解决目前现代信息检索在网络环境下所出现的一些问题，从而实现信息检索的智能化、个性化与标准化发展要求。欧盟早在 2002 年就开始研究基于网格的信息检索项目研究员。它在同年启动的 GRACE 计划的目标就是开发一个基于网格技术的支持实时数据、灵活数据分配和计算资源的信息检索系统。GRACE 系统是欧盟数据网格项目中的一个重要部分。随着网格技术研究在世界各地的兴起，基于网格的信息检索技术将成为未来的一个发展方向，用户能够以透明的方式获取资源，用户无须考虑资源的位置和获取时间，实现获取资源的一站式服务。

6. 专业化信息检索

现代信息检索技术的另一个发展趋势是检索专业化。专业化信息检索是指面向某一特定专业或学科领域，提供高质量的专业信息检索服务。专业化信息检索需求的出现主要因为网络信息资源越来越丰富，而综合性检索系统（比如搜索引擎）查询专业信息越来越困难，效率比较低，往往不能检索到高质量的专业信息。发展专业化检索将是未来的一个研究热点。专业化信息检索将只涉及某一学科、某一领域或特定需求的信息，这些信息相对集中，且其编制通常有本专业领域人员参与，因此它不仅可以提高检索速度，还可以提高信息专指度，加大检索深度和检索力度，从而提高查全率和查准率。目前在某些领域已经存在专业搜索引擎，而且这种数量必将越来越多。国际上著名的 PubMed 就是美国国家医学图书馆开发的医学专业信息的检索工具。世界范围内学科信息门户的兴起也是专业化信息检索的一种体现。英国资源发现网络（Resource Discovery Network）开发的社会科学信息门户（SOSIG）的宗旨就是为社会科学领域内的研究者提供筛选的高质量网络信息。中国科学院国家科学数字图书馆已建成包括物理和数学在内的六大学科信息门户，提供每一个学科领域内专业化的信息资源。另外，专业化信息检索不但体现在其搜索内容的专业性上，也体现在其搜索媒体性质的专门性上。比如，致力于检索图片的图片搜索引擎、致力于检索音乐的音乐检索系统，这种针对专门性质媒体的检索工具也在不断增多。专业化信息检索系统在提供专业信息方面有着大型综合检索引擎无法比拟的优势，它所采用的基本技术同综合引擎一样，而且基本上是成熟的技术，它们的发展没有技

术障碍,同时正符合了 Internet 发展的一个趋势:Internet 将更专业化、分工更细,因而专业化信息检索系统将能更好地为不同领域的用户提供个性化的服务。

总之,未来的信息检索发展将在理念、技术、人性化、智能化等方面取得全面突破,逐渐适应人脑的思维方式,实现智能、高效、快速而灵活的信息检索,最后达到随心所欲地查找、迅速获取所需信息的水平。当然,这些突破也需要计算机硬软件技术、通信技术、人工智能技术、可视化技术等相关技术支持,但无论如何,未来的信息检索一定会以一个崭新的面貌出现在人们面前,促进人们对无序信息世界的有序化组织,促进信息资源得到更为合理的查询、共享和利用。

3.3.2　信息检索的热点问题

(1) 智能检索或知识检索。传统的全文检索基于关键词匹配进行检索,往往存在查不全、查不准、检索质量不高的现象,特别是在网络信息时代,利用关键词匹配很难满足人们的检索要求。智能检索利用分词词典、同义词典,同音词典改善检索效果,比如用户查询"计算机"时,与"电脑"相关的信息也能检索出来;进一步还可在知识层面或者概念层面上辅助查询,通过主题词典、上下位词典、相关同级词典,形成一个知识体系或概念网络,给予用户智能知识提示,最终帮助用户获得最佳的检索结果。比如用户查询"计算机"时,用户可以进一步缩小查询范围至"微机"、"服务器"或扩大查询至"信息技术"或查询相关的"电子技术"、"软件"、"计算机应用"等范畴。另外,智能检索还包括歧义信息和检索处理,如"苹果",究竟指的是水果还是手机与平板电脑的品牌,"华人"与"中华人民共和国"的区分,将通过歧义知识描述库、全文索引、用户检索上下文分析以及用户相关性反馈等原理进行联合处理,高效、准确地反馈给用户最需要的信息。

(2) 知识挖掘。知识挖掘目前主要指文本挖掘,目的是帮助人们更好地发现、组织、表示信息,提取知识,满足信息检索的高层次需要。知识挖掘包括摘要、分类(聚类)和相似性检索等方面。

(3) 自动摘要。自动摘要就是利用计算机自动地从原始文献中提取文摘信息。在信息检索中,自动摘要技术有助于用户快速评价检索结果的相关程度。在信息服务中,自动摘要有助于多种形式的内容分发,如发往 PDA、手机等。相似性检索技术基于文档内容特征检索与其相似或相关的文档,是实现用户个性化相关反馈的基础,也可用于去重分析。自动分类可基于统计或规则,经过机器学习形成预定义分类树,再根据文档的内容特征将其归类。自动聚类则是根据文档内容的相关程度进行分组归并。自动分类(聚类)在信息自组织、智能导航方面非常重要。

（4）异构信息整合检索和全息检索。在信息检索分布化和网络化的趋势下，信息检索系统的开放性和集成性要求越来越高，需要能够检索和整合不同来源和结构的信息，这是异构信息检索技术发展的基点，包括支持各种格式化文件，如 TEXT、HTML、XML、RTF、MS Office、PDF、PS2/PS、Marc、ISO 2709 等格式化信息文档；支持多语种信息的检索；支持结构化数据、半结构化数据及非结构化数据的统一处理；与关系数据库检索的无缝集成以及其他开放检索接口的集成等。所谓"全息检索"的概念就是支持一切格式和方式的检索，从目前实践来讲，发展到异构信息整合检索的层面，基于自然语言理解的人机交互以及多媒体信息检索整合等方面尚有待取得进一步突破。另外，从工程实践角度，综合采用内存和外部存储的多级缓存、分布式群集和负载均衡技术也是信息检索技术发展的重要方面。随着互联网的普及和电子商务的迅猛发展，企业和个人可获取、需处理的信息量呈爆发式增长，而且其中绝大部分都是非结构化和半结构化数据。内容管理的重要性日益凸显，而信息检索作为内容管理的核心支撑技术，随着内容管理的发展和普及，亦将应用到各个领域，成为人们日常工作、学习与生活的密切伙伴。

（5）自然语言处理和问答系统。自然语言处理的应用，在一定程度上提高了信息检索的效果。例如互联网舆情分析系统引入主题检测和热点发现技术，对文本信息态度进行分析，为舆情监管和互联网信息挖掘的研究提供了数据积累和技术支持。此外，还有学者涉及词义消歧模型的研究等。问答系统是通过处理用户提出的自然语言问题，抽取有效信息，最后以自然语言给出答案的一个工具，它能够给用户提供相对简洁、准确的结果，因此越来越受到学者的关注。这方面的研究主要包括对开放域问答系统进行了综述，介绍了其系统框架、主要技术和评测方法；模式推理在问答系统中的应用以及模式推理的基本方法，实现了常量、变量一体化索引的算法，并给出了算法分析等。

（6）多媒体检索。多媒体检索技术是对图片、音乐、视频等媒体对象的检索处理，也是学者们当前关注的重要热点之一。例如一种基于小波和 Hough 变换的放射不变性商标检索方法；一种基于音频信息重复性的广告检测方法；能够对海量音频信息进行快速检索并找到检索词、发音准确位置的关键音检索系统；通过对足球比赛视频中的场地信息和运动信息的分析系统，提出有效分割场地和运动员的新方法。

（7）信息检索模型与算法。信息检索模型及算法是信息检索领域的核心，其涉及的高频主题词包括向量空间模型、聚类、算法、查询扩展、关联规则和机器学习，这方面的研究主要表现为对检索模型的改进与完善、扩展和应用。例如对面向信息检索的语言模型存在的数据稀疏问题，提出面向信息检索的近邻语言模型圈；解决排序学习中 pairwise 方法的问题，分别基于单层神经网络和双层神经网络的 RankNet 算法，加入 pointwise 损失

函数进行优化,并分别使用梯度下降算法和反向传播算法训练网络权重值,进而得到排序检索模型等。

(8) 文本分类、文本表示和信息安全。文本分类和文本表示在前两届全国信息检索技术学术会议中研究较多,相关研究也主要集中在对模型和方法的改进和完善上。例如通过对经典的 TF-IDF 函数和对信息特征选择方法的研究,提出了引入类间分布度、类内分布度和互信息因子的改进算法;新的特征选择和加权方法以类信息作为调节因子,使均匀分布于单个类中的特征更具代表性,弥补了传统文本分类方法的不足。

(9) 文本挖掘、信息抽取与信息过滤。文本数据挖掘(text mining)是指从文本数据中抽取有价值的信息和知识的处理技术。顾名思义,文本数据挖掘是从文本中进行数据挖掘(data mining)。文本挖掘种类有两类:即基于单文档的数据挖掘和基于文档集的数据挖掘。文本挖掘方法包括:文本分类(文本分类是一种典型的机器学习方法,一般分为训练和分类两个阶段)、文本聚类(文本聚类是一种典型的无监督式机器学习方法,聚类方法的选择取决于数据类型)、信息抽取、摘要和数据压缩。

信息抽取(information extraction,IE)是把文本里包含的信息进行结构化处理,变成表格一样的组织形式。输入信息抽取系统的是原始文本,输出的是固定格式的信息点。信息点从各种各样的文档中被抽取出来,然后以统一的形式集成在一起。这就是信息抽取的主要任务。信息以统一的形式集成在一起的好处是方便检查和比较。信息抽取技术并不试图全面理解整篇文档,只是对文档中包含相关信息的部分进行分析。

信息过滤是大规模内容处理的另一种典型应用。它是对陆续到达的信息进行过滤操作,将符合用户需求的信息保留,将不符合用户需求的信息过滤掉。通常可分为不良信息过滤和个性化信息过滤:不良信息过滤一般指过滤掉暴力、反动、色情等信息;个性化信息过滤类似于专业信息检索,帮助用户返回感兴趣的专业信息。

本 章 小 结

"信息检索技术"的概念含义可以从信息、检索与信息检索技术三个概念的递进与组合关系进行理解与把握。信息指的是事物的存在方式和运动状态,是对客观世界中各种事物的变化和特征的反映,是客观事物之间相互作用和联系的表征,是客观事物经过感知或认识后的再现。广义的信息检索是指将信息按一定的方式组织和存储起来,并根据用户的需要找出有关信息的过程,所以它的全称又叫信息存储与检索。信息检索技术是跨越多学科领域的信息组织与信息提取方法的融合技术,是针对信息获取(acquisition)、信

息表示（representation）、信息存储（storage）、信息组织（organization）和信息访问（access）的特有融合性技术。它不仅涵盖传统针对各种具体文献数据库的图书情报检索技术，也包括针对现代广域互联网的网络信息检索技术。

信息检索是一门多学科交叉的应用技术领域。自然语言处理、分布式计算、数据库技术、数据挖掘、情报学、社会学等多个领域的原理与方法，对信息检索技术研究的拓展与深入有重要帮助与促进作用。

包括可视化和多样化、集成化、个性化、智能化、基于网格和云计算的信息检索技术、专业化信息检索等热点研究领域，未来的信息检索技术将在理念、技术、人性化、智能化等方面取得全面突破，逐渐适应人脑的思维方式，实现智能、高效、快速而灵活的信息获取的目的。

对于信息检索的用户而言，信息检索通常要具备四个基础素养：用户信息意识、信息源掌握、信息获取能力、信息共享与利用。信息检索包括书目检索、数据检索、事实检索、全文检索、图像检索和音频检索等基本类型。

智能检索或知识检索、知识挖掘、自动摘要、异构信息整合检索技术和全息检索、自然语言处理和问答系统、多媒体检索技术、信息检索的模型与算法、文本分类、文本表示和信息安全、文本挖掘、信息抽取与信息过滤等各个层面的信息检索研究，日益成为目前国内外信息检索的主要研究热点。

本章思考与练习题

1. 信息检索的含义是什么?
2. 信息检索包括哪些方面的内容?
3. 世界上第一个信息检索系统产生在何时何地? 它有哪些基本信息服务功能?
4. 举例说明“非结构化信息”的含义。
5. 信息检索有哪些基本类型?
6. 信息检索有哪些支撑领域?
7. 说明信息检索的主要发展趋势。
8. 信息检索有哪些热点问题?

第4章　信息检索的方法与策略

4.1　信息源及其类型

信息源也就是我们在检索过程中经常接触到的不同信息集合或者不同检索对象实体，也就是我们获得原始信息内容的来源。关于信息源的分类，从不同的角度出发有不同的分类方法。例如，按信息内容的时效性特征，可分为消息性信息源、资料性信息源和知识性信息源；按反映信息的主客观性可将其分为客观信息源和主观信息源，或者分为事实信息源和分析信息源；按信息的学科内容可将其分为社科信息源、科技信息源、经济信息源、军事信息源、体育信息源等。

1. 依据信息内容的加工层次划分

（1）零次信息源。零次信息源指存在于或存储于非正规载体上未经任何加工处理的源信息类型，例如书信、论文手稿、笔记、实验过程记录、会议记录、演讲、口语交流等。这是一种零星的、分散的和无规则的信息源。这类信息源是近几十年来被图书情报学界、信息学界、社会学界等领域逐步认识和重视的信息对象与获取来源，它具有原始性、原创性、新颖性、分散性和非检索标识等非规范特征。

（2）一次信息源。一次信息源又称原始获得信息源，是指直接将理论、设计、试验、生产、研究等信息或知识成果经过整理后，记录在正式和规范物理载体上的信息源。一次信息源的载体形式丰富，也称原始文献信息对象。例如各类图书内容、专著原文、期刊论文、研究报告、会议论文、专利说明书、学位论文、技术标准等。一次信息源反映了人类科学、技术、社会、经济和文化发展的直接成就，是人类文明和财富的象征，它具有新颖性、创造性和系统性等特征。一次信息源是信息检索与获取的直接对象和主要内容，信息检索的直接目标就是查找所需的一次信息源。

（3）二次信息源。二次信息源指应用科学的信息处理技术和方法，将分散无序的一次信息源（例如将“汗牛充栋”的书籍等一次信息源处理为“图书目录检索库”）内容进行加工、整理，使之成为检索系统中有序的结构化信息。二次信息源的各种载体集合通常称为

信息检索工具,例如文摘、书目、索引、指南、搜索引擎等。二次信息源具有浓缩型、汇集性和有序性等特点,它是查找一次信息源的工具。大学生学习信息检索的工具与方法,主要是指学习二次信息源的检索与利用。

(4)三次信息源。三次信息源指对零次、一次和二次信息源进行综合分析并处理加工后的检索对象。三次信息源的内容包括述评、研究综述、进展报告、数据手册、年鉴、专业词典等。

各次信息源之间的关系。从零次文献、一次文献、二次文献到三次文献,是一个由分散到集中,由无序到有序,由博而精地对知识信息进行不同层次的加工过程。它们所含信息的质和量是不同的,对于改善人们的知识结构所起到的作用也不同。零次信息源是形成一次信息源的原始信息素材,一次信息源是形成二次信息源和三次信息源的基础,没有一次信息源就不会产生二次信息源和三次信息源。利用二次信息源检索和利用一次信息源,也可以进一步形成三次信息源,二次信息源是一次信息源的浓缩与检索应用工具,但三次信息源又在二次信息源中得到反映,所以二次信息源既是检索一次信息源的工具,又是检索三次信息源的工具。信息源加工层次分类及其关系如图 4-1 所示。

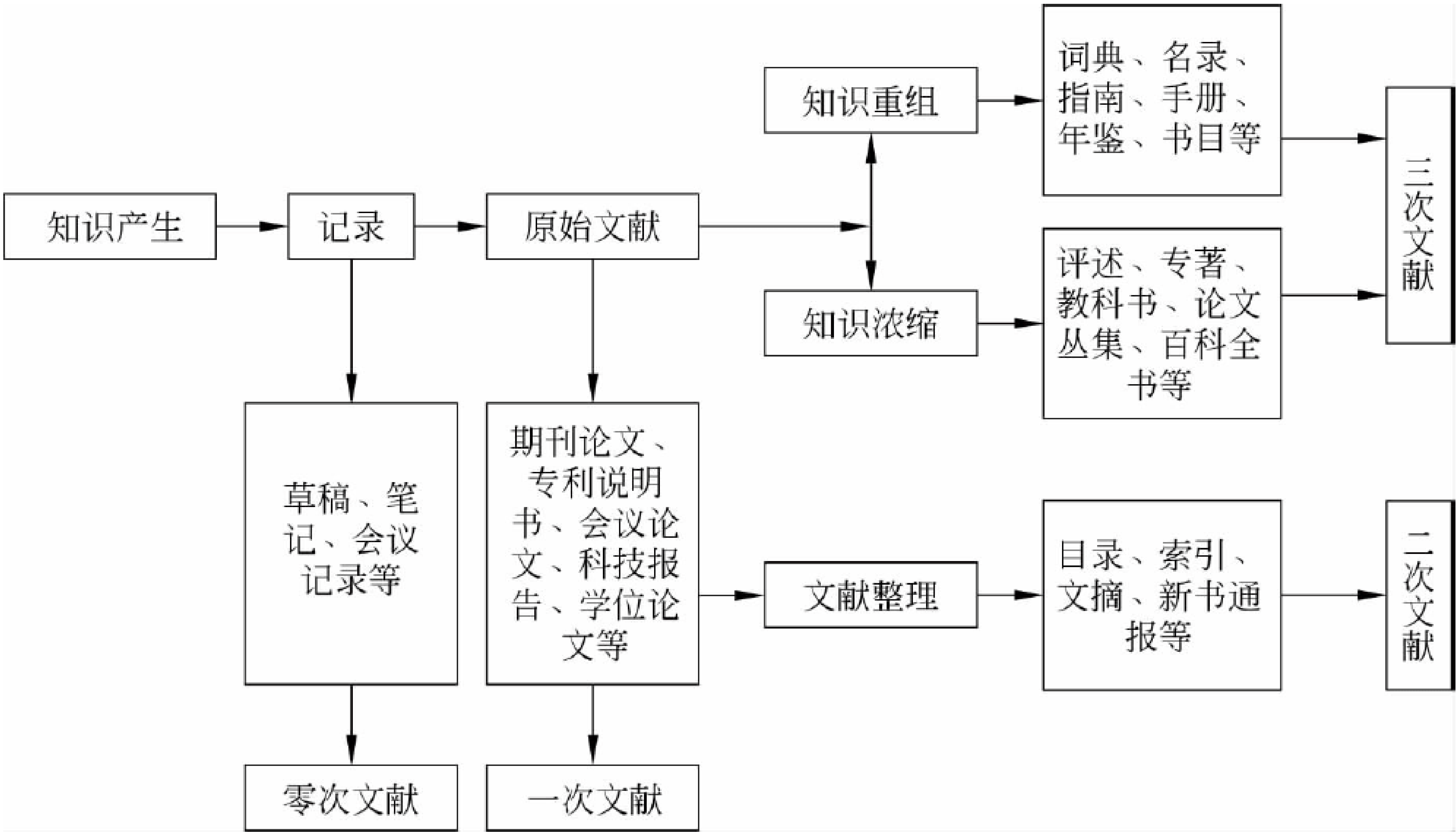


图 4-1 信息源加工层次分类及其关系图

2. 信息源载体的物理类型

信息源载体随着人类社会的进步而不断发展,其形式从古代的甲骨文到传统的纸质

载体，直到现代的电子型和网络磁性介质载体，经历了一个漫长的发展过程。现代信息载体的记录和存取技术已经飞速发展，使信息源载体形式进入了一个崭新的新时代。

（1）印刷型。印刷型指通过油印、铅印、胶印、喷墨、激光等各种印刷手段将信息记录在纸张上的信息检索资源，这是沿用了近千年的传统载体形式，是各类信息源载体的主体对象，也是检索与利用的重要主体。其特点是使用方便，易于阅读，但需要占用大量空间，不便于整理和保存。例如目前高校图书馆和社会图书馆都藏有海量的书籍，也是学生检索、获取与学习利用的主要对象。

（2）缩微型。缩微型指通过光学技术将印刷型信息和图像拍摄或复制在透明或不透明的感光材料载体上的文献。它又可分为缩微胶卷和缩微平片。其优点是体积小，易于保存，存储密度高，例如每张 $105\times148mm^2$ 的平片可容纳 3200 页的图书内容。其缺点是阅读不便，并需要专门的阅读设备和环境。目前这类信息源基本被淘汰，除了收藏与研究价值外不具有用户的普遍检索与利用特性。

（3）声像型。声像型又称直感型或试听型信息源，指通过专用设备，使用声、光、磁、电等技术将信息以声音、图像、影视和动画等形式表现出来，具有直观形象的优点。它在帮助人们观察罕见的自然现象、探索物质的微观结构或者辅助用户专业性知识学习等方面，能起到印刷型信息源不能具备的独特作用，其缺点是需要借助录放机、计算机、DVD 机、音箱和显示器等设备才能检索和利用声像型信息资源。

（4）磁介质型。磁介质型常常称为计算机型数字化信息资源，指通过编码、指令操作、程序设计与编程、数据库存储与管理、磁盘服务（包括分布式网络云盘）等技术融合，将信息转换成计算机终端机或服务器能够独立使用的数据，也可网络化大规模检索与共享的数字化的磁介质型信息资源。这是信息社会的信息存取主要手段与方法，它具有存储容量大、存取速度快、传播广泛，以及原记录可以及时修改、删除或更新等特点。

4.2　信息源的出版发行与共享类型

信息源的出版发行与共享类型，是依据信息源载体内容的性质、作用、出版发行方式、检索、利用与共享特点来对信息源进行辨认、识别、检索与利用，目的是让大学生认识这些信息源类型的不同作用，以提高信息检索与利用的针对性与目的性。

（1）图书。这是一种论章成册的出版物，是对已有研究成果、生产技术、实践经验或某一知识体系的论述或概括。它的基本素材来自期刊论文、会议论文、研究报告、学位论文等一次信息源和著者本人的研究和学术成果。由于经过著者或编者的选择、核对、鉴

别、提炼和加工，因而内容比较成熟、全面和系统，是传播知识、教育和培养人才的重要工具。图书的出版周期较长，报道速度较其他信息源要慢。

图书按其内容性质和作用可分为：①普及读物；②教科书；③丛书；④专著；⑤论文集；⑥参考工具书，如书目、索引、字典、词典、手册、年鉴、指南和百科全书等。

（2）期刊。期刊又称杂志，这是一种有固定名称，有一定出版规律，每期由多篇论文组成的连续出版物。其特点是出版周期短，报道速度快，数量大，内容丰富新颖，原创性高，能及时反映当代社会和科技的发展水平和动向，它所刊载的科学事实、数据、理论、技术、方法、构思和猜想，都是专门学习、科学研究的重要参考信息。因此，大学生和研究人员一般都要经常查阅期刊论文，借以了解动态，掌握进展，开阔思路，吸取他人的经验或思路，以改进或提高自身学习和工作效率。

期刊按其报道内容性质，可分为科普性期刊、技术性期刊、学术性期刊、信息性期刊、检索性期刊和数据性期刊等种类。

（3）报纸。报纸（newspaper）是以刊载新闻和时事评论为主的定期向公众发行的印刷出版物。这是出版周期最短、发行量最多的一种出版物，一般可分为综合类报纸和专业类报纸，前者报道内容极为广泛，涉及政治、经济、军事、科技、文化艺术、生活等方面，一般以消息性信息居多，后者侧重报道某一方面内容的消息性信息和知识性信息。报纸是大众传播的重要载体，具有反映和引导社会舆论的功能。报纸通常为散页印刷，不装订、没有封面，但有固定名称，面向公众定期、连续发行。现代报纸每日出版一次，称为日刊；或者每周出版一次，称为周报。

报纸的优点是：可随时阅读，不受时间限制；互相传阅，读者人数可以是印刷数的几倍；即使阅读或理解能力较低的人，亦可相应吸收报章的信息；由于互联网的崛起，网上版报纸比传统印刷版的信息量要大得多、传播速度要快得多、受众也更加广泛，同时传统报业公司都纷纷建立了自己的在线报纸门户网站。

（4）会议文献。会议文献是指在各种综合性、行业性或专业性会议上发表的论文和报告。此类信息的学术性和专业性都较强，信息的专业性价值较高。往往反映了当前的学科进展或行业发展动态，是获取最新信息的重要来源。

会议文献按其出版形式又可分为：①连续出版物，以定期或不定期的形式连续出版，一般以会议录居多；②图书类，以会议名称作为书名或另加专门书名，按图书出版发行，一般以会议论文集居多；③期刊类，将会议论文刊载在某一期的期刊上，以专刊或增刊的形式出版发行。了解和掌握各种会议论文的不同出版形式和收藏特点，对于索取“原始信息全文”具有实用意义。

(5) 科技报告。科技报告是科技人员或科技研究机构从事某一专题研究所取得的成果和进展的实际记录。其特点是反映新技术、新学科和新知识,报道信息速度较快,内容比较专深、新颖,数据比较可靠,保密性较强,有相当一部分科技报告资料不公开发行。科技报告每份单独成册,有专门编号,用以识别报告类型及其主持机构。

(6) 专利文献。专利信息是与专利制度有关的所有专利文件与技术资料的总称,包括专利说明书、专利公报、专利分类表、专利检索工具以及专利的法律文件。其中主体是专利说明书,它具有统一编号、数量大、内容丰富、实用、可靠、新颖、原创以及报道迅速等特点。专利信息总体上分为四类:专利规范信息、发明专利信息、外观设计专利信息和实用新型专利信息。在信息的价值与使用价值方面,专利中的发明专利信息最高。

(7) 学位论文。学位论文是指高等院校或研究机构的毕业生和研究生为取得各级学位而撰写的学术论文,它按级别可分为学士学位论文、硕士学位论文和博士学位论文。其中研究生论文(尤其是博士学位论文)带有一定的创造性,所论及的内容一般比较专深,对科研、生产和教学有较大的参考价值。作为大学生,不仅要充分了解、查询和利用学位论文资源,而且也是学位论文的直接生产者。

(8) 技术标准。技术标准是指描述有关产品和工程质量、规格、工艺流程及其测试方法等的技术文件。技术标准是一种经权威机构批准的规章性信息资源,具有一定的法律约束力。按其约束力可分为法定标准、推荐标准和试行标准;按其执行范围可分为国际标准、区域标准、国家标准、专业标准和企业标准等;按其内容可分为基础标准(包括术语、符号、单位、定义等)、产品特性标准(包括特性、尺寸、形状、成分、质量等)以及方法标准(包括生产方法、作业方法、试验及检测方法等)。

(9) 政府出版物。政府用以发布政令和体现其思想、意志、行为的物质载体,同时也是政府的思想、意志、行为产生社会效应的主要传播媒介。政府出版物是指各国政府及其所属分支机构所发表的各类公务性、政策性等行政类文件和科技型文件。政府出版物数量巨大,内容广泛,出版迅速,资料可靠,是重要的信息源。

(10) 产品样本资料和说明书。产品样本资料是指厂商或贸易机构为宣传和推销其产品而印发的免费赠给消费者的资料。如产品目录、产品样本、产品说明书、产品总览、产品手册等。它们大多是对定型产品的性能、构造原理、用途、使用方法、操作规程、产品规格等所做的具体说明。产品样本资料图文并茂,形象直观,所反映的技术较为成熟,数据较为可靠,对技术革新、选型、设计、试制新产品以及引进设备等均有一定的参考价值。产品样本资料随着产品的更新换代而更新,而且有一部分产品是试销或试验性产品,在查询与利用该类信息时应予以注意。

(11) 技术档案。它是在生产建设过程和技术研发活动中形成的具体工程对象的技术文件、图样、图表、照片、原始记录或其复制品。其内容包括任务书、审批文件、研究计划、技术指标、技术措施、调查材料、设计计算、工艺记录、研究结论等信息。它是科研和生产建设中积累经验、提高质量的重要依据。此类信息资源具有明显的保密性和内部控制使用的特点。

以上所述的11种出版类型中,图书、期刊、会议文献、科技报告、专利文献、学位论文和技术标准均有其相应的二次信息源,即检索工具。所以查找起来比较方便,而政府出版物、产品样品和说明书、技术档案、报纸则多数没有相应的二次文献,所以查找极为不便。

近年来,有关部门对政府报告、技术档案、重要报纸等组织专门人员制作相应的二次文献,解决了检索与获取不便的问题。

4.3 信息源类型的辨别

信息检索的目的就是从"信息海洋"中查找出不同类型的信息,以满足不同的信息需要。为此,首先必须鉴别信息源的不同类型,以便按此获取相应的原始信息全文。当我们检索多种类型的检索刊物或者利用各类论文后所附录的参考文献来扩大检索线索与范围时,都会遇到识别信息源类型的问题。各种检索工具(或检索数据库)所汇聚的各类信息,由于普遍采用规范化的著录格式,有明显的文献类型标识,所以识别时并不困难。但科技专著、技术报告和各类论文后所附录的参考信息源,除中文期刊的参考文献目前已统一采用标准著录格式而不难识别外,尤其是外文出版物所附录的参考信息源的著录格式没有统一的标准,特别是文献的出处项,其正斜体、大小写、简称(缩写)等项目因国、因人、因出版物而异,甚至五花八门,加之有些参考文献的著录项目不全,没有明显的信息源类型标识,所以增加了识别的难度。因此,学习和掌握识别信息源类型的一些基本原则和方法具有一定的实用意义。

1. 图书

图书除了著(编)者和书名之外,识别其信息源类型的明显标识是出版单位名称、出版地及出版时间。

【例 4-1】 C. Koelbel, D. Loveman, R. Schreiber, G. Steele, Jr., and M. Zosel①, *The High Performance Fortran Handbook*②. Cambridge, MA③: MIT Press④, 1994⑤.

【例 4-2】 C. P. Wong, J. M. Segelken, and C. N. Robinson⑥, "Chip on board encapsulation,"⑦ In Chip on Board Technologies for Multichip Modules②, J. H. Lau①,

Ed. ,Yew York[③]：Van Nostran Reinhold[④],1994[⑤],pp. 470-503[⑧].

例中,①图书著(编)者；②书名(一般用斜体)；③出版地；④出版社名称；⑤出版时间；⑥著者；⑦论文题名；⑧起止页码。

上述两例中,例 4-1 较易识别,因它有较明显的出版社标识“MIT Press”(Press 即出版社)。例 4-2 不但要识别出出版社名称(Van Nostran Reinhold),而且还应识别出书名、编者和论文著者。这是一本以图书形式出版的专题论文集,如果将编者和书名误认为论文著者和论文题名,那么在馆藏书目数据库中就无法检索到该书。此外,有些丛书编有卷号或期号,不能与期刊信息相混淆。

2. 期刊论文

期刊论文的出处项一般包括刊名、卷、期、页码及出版时间。其明显的标识是有卷、期号和起止页码。

【例 4-3】 D. E. Everitt and N. W. Macfadyen[①],“*Analysis of multicellular mobile radio telephone systems with loss*,”[②] Br. Telecom Tech. J. [③], Vol. 1, no. 2[④], pp. 37-45[⑤],1993[⑥].

【例 4-4】 Y. Yang, G. M. Masson[①], Broadcast ring sandwich networks[②], *IEEE Transactions on Computers*[③]44(10)[④](1995)[⑥]1169-1180[⑤].

例中,①论文著者；②论文题名；③刊名(外文刊名有时为斜体)；④卷、期号；⑤起止页码；⑥出版日期。

上述例子的刊名、卷期号均有明显的标识,不难识别。例 4-4 的刊名、卷期号、出版日期和起止页码采用与例 4-3 不同的著录方式,识别时应注意判断。

3. 会议论文

会议信息源的出处项包括会议或会议录名称、出版时间和页码,有的还有会议地址或主办单位。其特点是有反映会议信息的明显标识,如 Proceedings、Symposium、Meeting、Workshop、Colloquium、Convention 等。

【例 4-5】 Soumyanath and J. Von Arx[①],“An anolog parallel processor for the dynamic programming paradiam”[②], *Fifth Ann. IEEE Int. ASIC Conf. Exhibit*[③], Rochester NY[④],1992[⑤],pp. 557-560[⑥].

【例 4-6】 R . Mendis, M. T. Bishop, and J. F. Witte[①],“Investigations of voltage flicker in electric arc furnance power systems”[②], Proc. *IEEE-IAS Annu. Meeting*[③], 1994[⑤],vol. 3[⑦],pp. 2317-2325[⑥].

例中,①论文著者；②论文题名；③会议名称或会议录名称(外文刊名多数用缩写斜

体表示)；④会议地址；⑤会议时间或会议录出版时间；⑥起止页码；⑦会议录卷号。

上述两例的著录方式略有不同，其中例 4-5 中③明确给出了会议名称及开会地址，例 4-6 中③则为会议录名称(Proc. 是 Proceedings 的简写，即会议录的意思)并有卷号，识别时勿误认为期刊论文。

4. 技术报告

技术报告的明显标识是其信息源出处项标有“Report”，并列出相应的报告号，有时还有合同号(Contract)、入藏号(Accession)及出版机构等。

【例 4-7】 Butler, R and Lusk, E①, User's Guide to the p4 parallel programming system②, Technical Report ANL-92/17③, Argonne National Laboratory USA④ (October 1992)⑤.

【例 4-8】 “Computer assisted drawing information capture”②, Report NP-7179-CCML③, Vol. 1⑥, Electric Power Research Institute, Palo Alto, CA④, Jan. 1991⑤.

例中，①著者；②报告题名；③报告号；④机构名称及所在地；⑤报告发表日期；⑥报告卷号及此篇报告分若干卷，绝大多数报告无此标识。

5. 专利

识别专利信息源的主要依据是其出处项有专利(Patent)国别代号和专利号，有时还列出申请号(Application)。

【例 4-9】 T. A. D. Riley①, “Frequency syntheszers”②, U. S. Patent 4965531③, Oct. 1990④.

例中，①专利权人；②专利说明书题名；③专利国别代号及专利号；④专利批准日期。

6. 学位论文

学位论文的主要特点是其信息出处项一般有 Ph. D. thesis, Ph. D. dissertation, Master's thesis, M. S. thesis 等标识，thesis 为学位研究论文，Ph. D. 则明显表示具体学位。

【例 4-10】 G. T . Byrd①, “Communication mechanisms in distributed shared memory multi-processors”②, Ph. D. dissertation③, Stanford Univ. , Stanford, CA④, Aug. 1998⑤.

【例 4-11】 H. Hauson①, “Connection management functions of a private wireless ATM network”②, Master's thesis③, Helsink Univ. Technology④, Mar. 13, 1996⑤.

例中，①著者；②论文题名；③论文类型；④著者所在单位及地址；⑤论文发表日期。

7. 技术标准

技术标准的识别主要依据是其出处项一般均有 Standard(Std),Specification 以及标准颁发单位及标准代号,如 ISO,NBS,ANSI,CCITT,GB 等。

【例 4-12】 IEEE Guide for Harmonic Control and Reactive Compensation of State Power Conventers①,IEEE Std. ⑤19②,1981(updated 1989)③.

【例 4-13】 MPEG Requirements Group④,"MPEG-4 requirements document V. 4"①,ISO/IEC JTCI/SC29/WG11 N1727②,July 1997③.

例中,①标准名称;②标准制订机构代号及其标准号;③标准公布日期(例 12 中括号内为修改日期);④标准机构下属专业组。

8. 其他文献

除以上七种主要外文信息源识别外,在参考文献中还经常出现其他各种类型的信息源,如数据手册、技术说明书、内部文件、私人信函等,有的参考信息源识别十分困难,我们应根据文献题名及出处的某些特征进行仔细分析和判断。

【例 4-14】 MCNC① Open Architecture Silicon Implementation Software User's Manual②,Microelectronics Corporation of North Carolina,USA③(1990)④.

【例 4-15】 Private email Correspondance⑤,T. G. Mattson⑥ of Intel Supercomputer Division,USA⑦(email address: tgm @ SSD. Intel. Com-)⑧(1994)④.

例中,①机构缩写名;②文献题名;③机构全称及所在国名;④出版(信息生成)日期;⑤文献类型,这是一篇私人 E-mail 通信;⑥发信者;⑦发信者所在机构及国别;⑧E-mail 地址。

4.4 检 索 工 具

4.4.1 检索工具的基本功能

信息检索工具是以压缩形式存储、报道和查找信息线索或原始信息全文的工具,它是经过对信息进行搜索整理、特征分析和组织加工后的产物,同时又是信息检索的主要手段和条件。它包括传统的检索工具,例如科学引文索引 SCI;也包括网络检索工具例如 Baidu 等。信息检索工具的主要功能表现在存储和检索两个方面。一方面将信息的外部特征和内容特征著录成多个可用的信息线索,并按照科学的体系和检索方法将信息检索项有序地组织起来,即信息特征的存储过程,这就是设计和编制检索工具的过程。另一方面,检索工具提供多种检索手段,使人们能够按照一定的检索方法和途径获得所需信息的

线索或原始信息全文,即检索过程,也就是利用检索工具获得所需信息的过程。信息在检索工具中的存储与检索过程简图如图 4-2 所示。

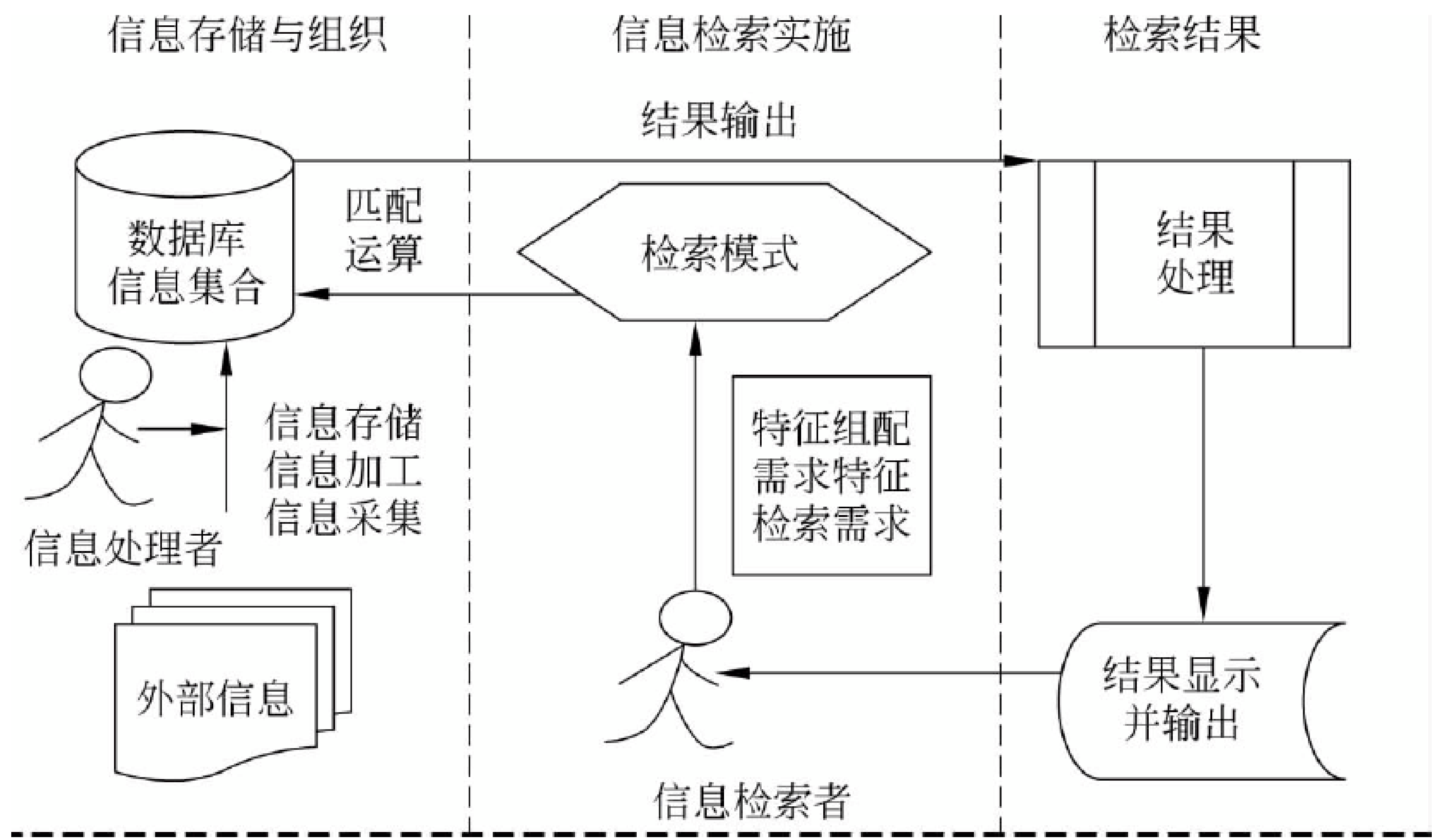


图 4-2　信息在检索工具中的存储与检索过程简图

信息报道及时、全面,存储规范、有序以及检索迅速和准确是对检索工具的基本要求,因此信息检索工具包含下列基本功能。

(1) 报道功能。检索工具以压缩的形式简明地揭示了信息的外部特征(例如书名、刊名、著者、号码、网页标题或网页链接地址等)和内容特征(例如标题、主题、摘要、分类、关键词等),供信息用户按照这些报道线索查找所需的原始信息。

(2) 标识功能。检索工具将所选择收录和分析整理后的信息按照一定的科学体系组织成一个有机的整体,同时进行多种检索标识(例如序号、代码号、主题词、关键词、学科类目等标识)。多种检索标识是系统(包括手工检索系统和计算机检索系统)标引人员和信息用户所共同遵循并进行彼此沟通的,这些"共同语言"标识也是提高检索工具的存储质量和检索与利用效率的重要基础。

(3) 辅助检索功能。为使信息用户能通过多种检索方法和途径获取信息,检索工具必须提供多种辅助检索手段,例如分类索引、主题索引、代码索引和著者索引以及机构索引等。辅助检索功能的完善程度不仅是检索工具的主要质量指标,而且也是影响信息用户能否充分实现信息资源共享的一个关键因素。在电子信息数据库或网络搜索引擎中,辅助检索功能十分强大,例如从时间、类型、学科范围、用户点击量等方面进行信息约束或限定,就起到了很好的辅助作用。

4.4.2　检索工具的类型

由于信息检索工具的著录性质、报道范围、载体形式和检索手段等特征的不同，检索工具有多种划分方法，通常按著录信息的特征进行分类。

1. 目录(content)

以某一高校的图书馆图书目录数据库检索界面为例(如图 4-3 所示)，在图书目录实例中，左侧的分类目录工具十分明显。通过目录工具(一级目录、二级目录或多级目录工具)，可以快速定位用户信息需求的范围，引导用户渐进式、深入细化并准确查询目标信息内容。



图 4-3　某图书馆图书目录数据库检索界面实例图

目录就是依据信息的外部特征为著录依据，记录具体信息生成或出版事项及其信息收藏的报道性工具，目的在于信息或知识的定向查找和集成发现。按组织形式可划分为国家书目、联合目录、馆藏目录、报刊目录、联机性和网络性目录库等多种类型。按信息组织和报道范围可将目录划分为专题性目录(例如图书目录、期刊目录等)和综合性目录(例

如跨库联合目录)。

2. 索引(index)

索引是对数据库表中一列或多列的值进行排序的一种结构。使用索引可快速访问数据库表中的特定信息。在数据库中,索引是一种与表有关的数据库结构,它可以使对应于表的结构化查询语句执行得更快。索引的作用相当于图书的目录,可以根据目录中的页码快速找到所需的内容。当表中有大量记录时,若要对表进行查询,第一种搜索信息方式是全表搜索,就是将所有记录一一取出,并与查询条件进行一一对比,然后返回满足条件的信息记录,这样做会消耗大量数据库系统运行时间,并造成大量磁盘输入与输出操作;第二种就是在表中建立索引,然后在索引中找到符合查询条件的索引值,最后通过保存在索引中的地址快速找到表中对应的记录。索引的快速构建结构如图 4-4 所示。

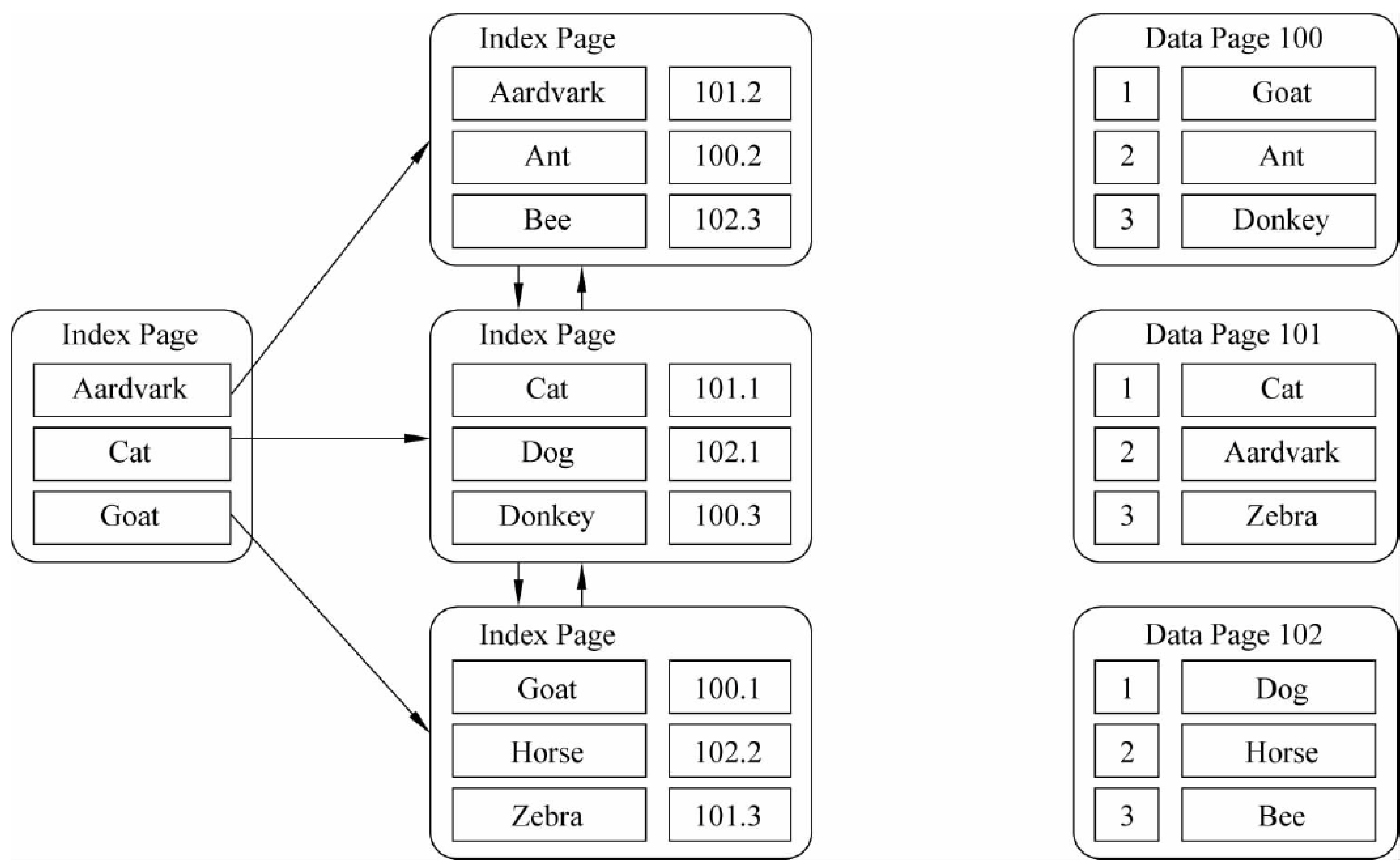


图 4-4 索引的快速构建结构示意图

索引工具就是将信息的一些外部特征或内部特征(例如题名、著者、主题、分类等)作为著录依据,并依此线索并引导出所需原始信息内容的检索工具。索引与目录的区别在于它不仅能揭示信息的外部特征,而且也能揭示信息的内部特征,例如主题索引、分类索

引、关键词索引等。索引既可单独出版，也可以附录形式出版，或者开发为网络数据库。大型的索引工具有“科学引文索引”(SCI)、工程索引(EI)、世界专利索引(WPI)等。

3. 文摘(abstract)

文摘在著录信息外部特征的基础上，还增加有揭示内容特征的摘要部分，它是系统地报道、积累和检索信息的主要工具，是最核心的检索工具。文摘型检索工具是将大量分散的信息全文，选择全文中重要的部分，以简练的形式组织为摘要，并按一定的方法组织排列起来的检索工具。在实践上，不仅传统的文摘检索工具有揭示信息内容的摘要内容，而且目前网络搜索引擎和大多数网络数据库都应用了揭示信息内容的摘要形式。依据文摘揭示信息内容的深度，它可分为指示性文摘和报道性文摘。指示性文摘：就是用简洁的语言简单说明信息的主题内容，以对文献题名做简要补充。报道性文摘：就是简要描述信息的主题内容，大多描述得较全面，一般包括主要内容、论点、结论、数据和图表等方面的内容。以某一高校的图书馆图书摘要数据库检索为例，以任意词“机器人”检索图书获得其中一本图书的摘要信息如下(如图 4-5 所示)：不仅包括图书内容摘要，也包括复本数与累借数、藏书地点和位置、作者与分类号、页数与价格等丰富的摘要性信息。

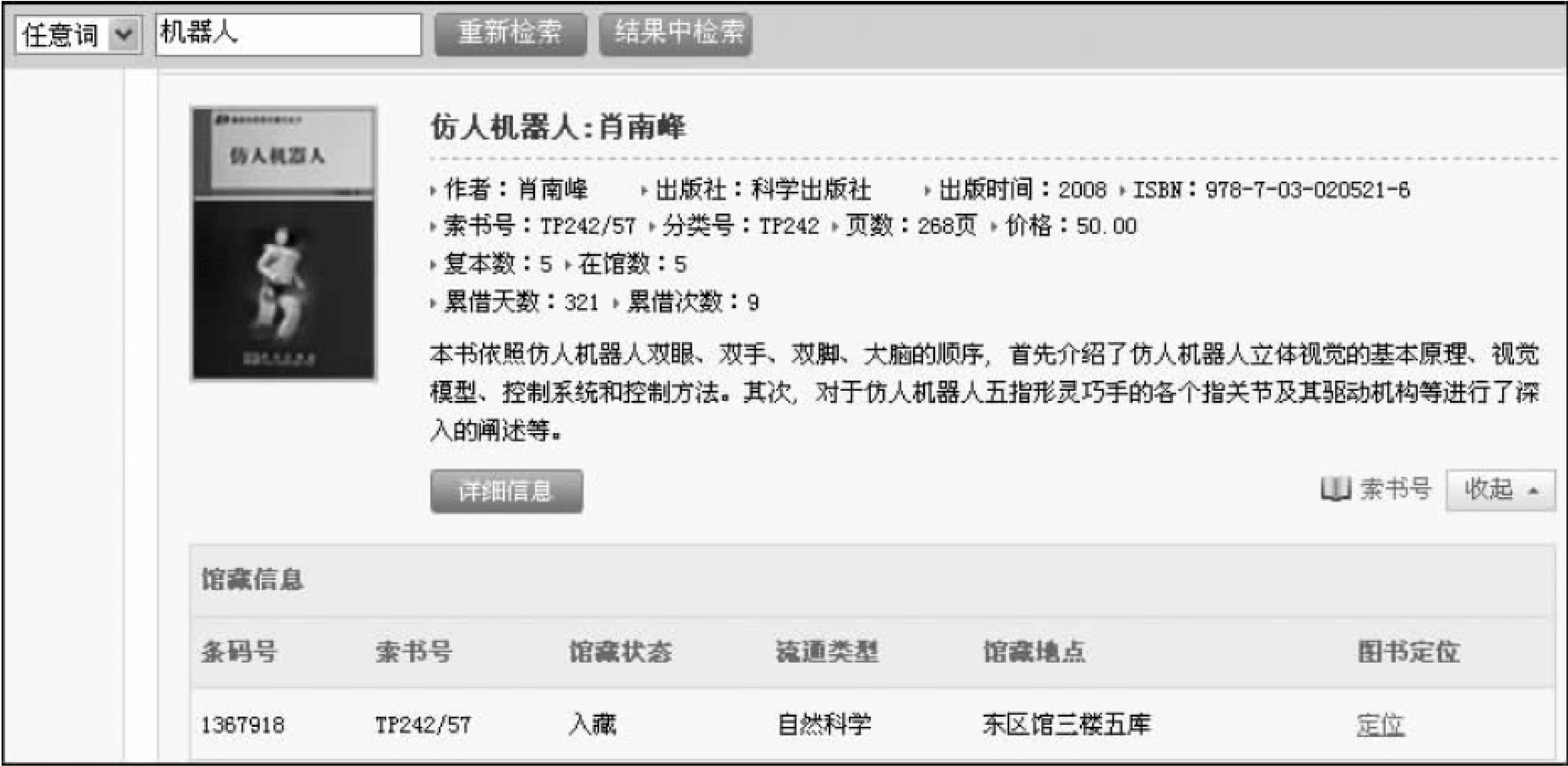


图 4-5　某图书馆图书摘要数据库检索为例

4. 参考工具书(reference)

参考工具书就是分析和著录大量具体而常用的科学数据与事实、以备查用的各种常用工具书的总称。例如，查找物理量、物质特性、经济统计数据、专业术语的含义、人物或公司名录、字词含义等大量自然科学和社会科学数据的检索工具。这类参考工具书包括百科全书、年鉴、手册、指南、名录、字词典等。图 4-6 是在英语词典“金山词霸”中，输入

“virtual reality”词语后的检索结果：结果中不仅有标准发音和中文含义，而且包含基础释义、双语释义、网络释义和行业释义等丰富的解析内容。



图 4-6 英语词典“金山词霸”的检索实例图

5. 搜索引擎(searching engine)

以计算机技术和通信技术为骨干的网络化环境，支撑了规模庞大而又纷繁复杂的网络信息资源，以检索网络信息为对象的主要检索工具——搜索引擎，在 20 世纪 80 年代末快速发展起来。搜索引擎就是将网络信息按一定分类方法组织起来，引擎软件自动通过搜索网址(也称域名或 IP 地址)的方式来实现网页信息的智能抓取、自适应组织与管理。网络信息检索的形式多样，既可以是一般信息内容线索(例如虚拟图书馆的书目在线检索)，也可以是原始全文信息检索；既可以是一般文本信息检索，也可以是图、文、声、像、动画、视频等多媒体信息检索。目前网络搜索引擎很多，例如 Baidu、Google、Yahoo、Gopher、Infoseek、Lycos、Archie、Goyoyo、Chinavigator 等。图 4-7 是搜索引擎 Lycos 的检索界面，它和著名的 Baidu、Google 等搜索引擎一样，用户检索界面简洁而明了。

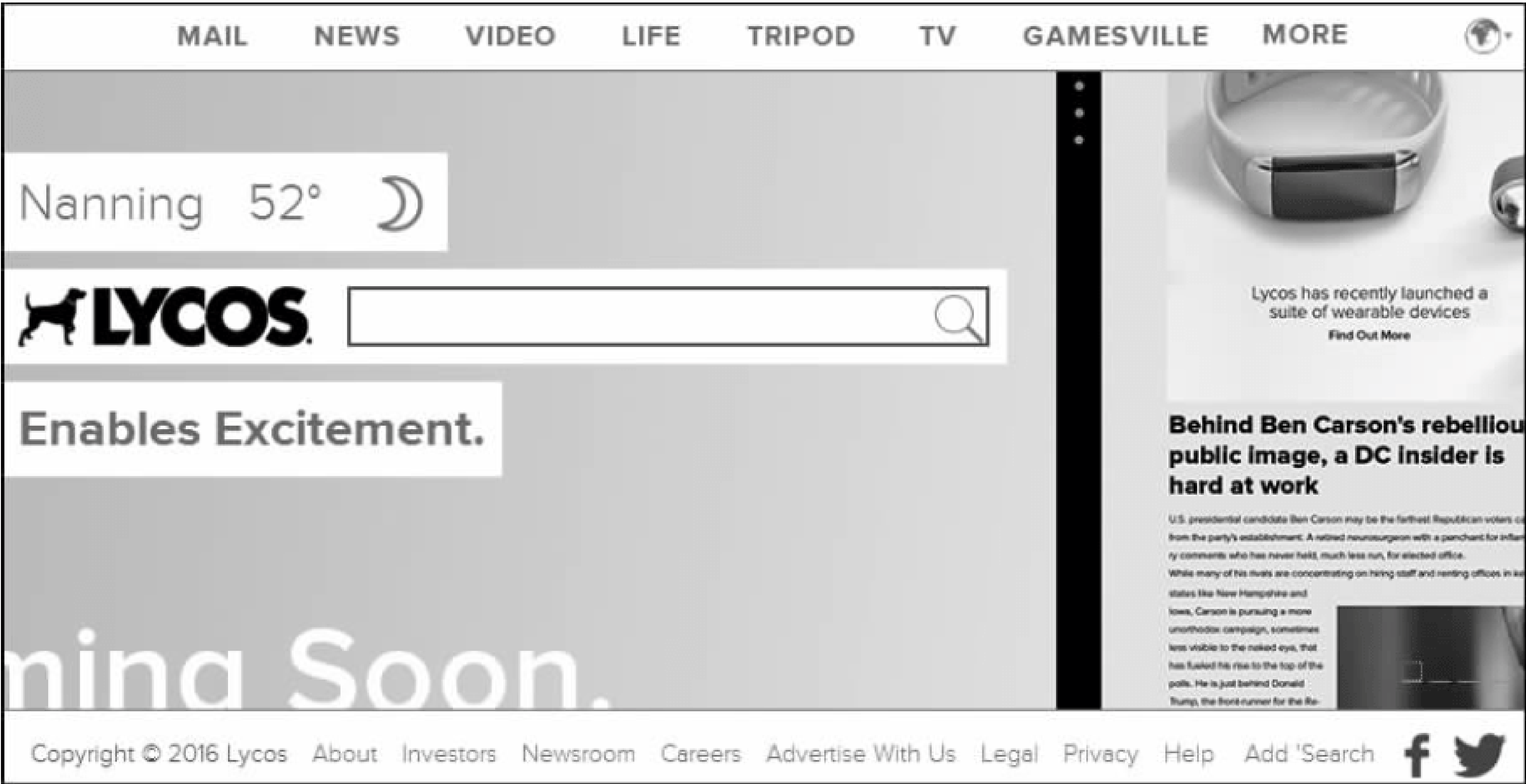


图 4-7　搜索引擎 Lycos 的检索界面

4.5　信息检索途径

信息检索途径多种多样，其中表明信息外部特征的相关途径有标题、责任者（或作者与发布者）、产生机构、序号、信息来源、产生时间、范围、路径、点击量或访问量等途径；与信息内容特征相关的途径有学科分类、主题和关键词、内容代码（例如化学分子式、图像色彩等）等途径。我国著名的“维普期刊数据库”检索途径实例如图 4-8 所示。



图 4-8　信息检索途径实例图

用户在信息检索与获取过程中常用的检索途径有内容分类途径、关键词或主题途径、著者途径、题名或标题途径、引文途径、序号途径、代码途径等。表 4-1 为检索途径分类表。

表 4-1 检索途径分类表

	信息特征	信息特征划分	信息检索途径
信息检索途径类型	外部特征	标题名 著者名 来源机构名 信息编号 载体类型编号 点击量与访问量 …	标题 著者 号码 来源 载体类型 点击量 …
	内容特征	分类范畴 主题词 关键词 其他(分子式等) …	分类号 主题词 关键词 分子式 …

1. 分类检索途径

分类途径是一种按学科分类体系来采集、存储和检索信息的途径。这一途径是以知识体系为中心进行分类组织的,能够体现信息内容的学科系统性,反映学科与信息内容的隶属、派生与平行关系,便于人们从熟悉的学科所属范围来查找所需信息,并且可以起到“触类旁通”的作用(例如同类信息或跨学科信息的查询)。分类检索途径使用分类语言、分类目录及分类索引等检索工具。例如,《科学文摘》的正文就是按照分类编排的,可以利用分类表,按分类进行查找。

信息和知识的体系化分类,比较典型且成熟的是我国的“中国图书馆分类法”,其分类严密、科学、完整且系统化(如下小体字所示),对大型检索系统或特定学科范畴的信息查询有很好的帮助作用。

A 马克思、列宁、毛泽东、邓小平理论

A1 马克思、恩格斯著作

A2 列宁著作

A3 斯大林著作

A4 毛泽东著作

A49 邓小平著作

A5 马克思、恩格斯、列宁、斯大林、毛泽东、邓小平著作汇编

A7 马克思、恩格斯、列宁、斯大林、毛泽东、邓小平生平和传记

A8 马克思主义、列宁主义、毛泽东思想、邓小平理论的学习和研究

B 哲学、宗教

B0 哲学理论

B1 世界哲学

B2 中国哲学

B3 亚洲哲学

B4 非洲哲学

B5 欧洲哲学

B6 大洋洲学

B7 美洲哲学

B80 思维科学

B81 逻辑学(论理学)

B82 伦理学(道德哲学)

B83 美学

B84 心理学

B9 宗教

C 社会科学总论

C0 社会科学理论与方法论

C 社会科学总论

C1 社会科学概况、现状、进展

C2 社会科学机构、团体、会议

C3 社会科学研究方法

C4 社会科学教育与普及

C5 社会科学丛书、文集、连续性出版物

C6 社会科学参考工具书

C7 社会科学文献检索工具书

C79 非书资料、视听资料

C8 统计学

C91 社会学

C92 人口学

C93 管理学

C94 系统科学

C95 民族学、文化人类学

C96 人才学

C97 劳动科学

D 政治、法律

D0 政治学、政治理论

D1 国际共产主义运动

D2 中国共产党

D33/37 各国共产党

D4 工人、农民、青年、妇女运动与组织

D5 世界政治

D6 中国政治

D73/77 各国政治

D8 外交、国际关系

D9 法律

E 军事

E0 军事理论

E1 世界军事

E2 中国军事

E3/7 各国军事

E8 战略学、战役学、战术学

E9 军事技术

E99 军事地形学、军事地理学

F 经济

F0 经济学

F1 世界各国经济概况、经济史、经济地理

F2 经济管理

F3 农业经济

F4 工业经济

F49 信息产业经济

F5 交通运输经济

F59 旅游经济

F6 邮电通信经济

F7 贸易经济

F8 财政、金融

G 文化、科学、教育、体育

G0 文化理论
G1 世界各国文化与文化事业
G2 信息与知识传播
G3 科学、科学研究
G4 教育
G8 体育
H 语言、文字
H0 语言学
H1 汉语
H2 中国少数民族语言
H3 常用外国语
H4 汉藏语系
H5 阿尔泰语系(突厥一蒙古一通古斯语系)
H61 南亚语系(澳斯特罗一亚细亚语系)
H62 南印语系(达罗毗荼语系、德拉维达语系)
H63 南岛语系(马来-波利西亚语系)
H64 东北亚诸语言
H65 高加索语系(伊比利亚一高加索语系)
H66 乌拉尔语系(芬兰一乌戈尔语系)
H67 闪一含语系(阿非罗一亚细亚语系)
H7 印欧语系
H81 非洲诸语言
H83 美洲诸语言
H84 大洋洲诸语言
H9 国际辅助语
I 文学
I0 文学理论
I1 世界文学
I2 中国文学
I3/7 各国文学
J 艺术
J0 艺术理论
J1 世界各国艺术概况
J19 专题艺术与现代边缘艺术
J2 绘画
J29 书法、篆刻
J3 雕塑
J4 摄影艺术
J5 工艺美术
[J59]建筑艺术
J6 音乐
J7 舞蹈
J8 戏剧、曲艺、杂技艺术
J9 电影、电视艺术
K 历史、地理
K0 史学理论
K1 世界史
K2 中国史
K3 亚洲史
K4 非洲史
K5 欧洲史
K6 大洋洲史
K7 美洲史
K81 传记
K85 文物考古
K89 风俗习惯
K9 地理
N 自然科学总论
N0 自然科学理论与方法论
N1 自然科学概况、现状、进展
N2 自然科学机构、团体、会议
N3 自然科学研究方法
N4 自然科学教育与普及
N5 自然科学丛书、文集、连续性出版物
N6 自然科学参考工具书
[N7]自然科学文献检索工具
N79 非书资料、视听资料
N8 自然科学调查、考察

N91 自然研究、自然历史
N93 非线性科学
N94 系统科学
[N99]情报学、情报工作
O 数理科学和化学
O1 数学
O3 力学
O4 物理学
O6 化学
O7 晶体学
P 天文学、地球科学
P1 天文学
P2 测绘学
P3 地球物理学
P4 大气科学(气象学)
P5 地质学
P7 海洋学
P9 自然地理学
Q 生物科学
Q1 普通生物学
Q2 细胞生物学
Q3 遗传学
Q4 生理学
Q5 生物化学
Q6 生物物理学
Q7 分子生物学
Q81 生物工程学(生物技术)
[Q89]环境生物学
Q91 古生物学
Q93 微生物学
Q94 植物学
Q95 动物学
Q96 昆虫学
Q98 人类学
R 医药、卫生
R1 预防医学、卫生学
R2 中国医学
R3 基础医学
R4 临床医学
R5 内科学
R6 外科学
R71 妇产科学
R72 儿科学
R73 肿瘤学
R74 神经病学与精神病学
R75 皮肤病学与性病学
R76 耳鼻咽喉科学
R77 眼科学
R78 口腔科学
R79 外国民族医学
R8 特种医学
R9 药学
S 农业科学
S1 农业基础科学
S2 农业工程
S3 农学(农艺学)
S4 植物保护
S5 农作物
S6 园艺
S7 林业
S8 畜牧、动物医学、狩猎、蚕、蜂
S9 水产、渔业
T 工业技术
TB 一般工业技术
TD 矿业工程
TE 石油、天然气工业
TF 冶金工业
TG 金属学与金属工艺

TH 机械、仪表工业
TJ 武器工业
TK 能源与动力工程
TL 原子能技术
TM 电工技术
TN 电子技术、通信技术
TP 自动化技术、计算机技术
TQ 化学工业
TS 轻工业、手工业、生活服务业
TU 建筑科学
TV 水利工程
U 交通运输
U1 综合运输
U2 铁路运输
U4 公路运输
U6 水路运输
[U8]航空运输
V 航空、航天
V1 航空、航天技术的研究与探索
V2 航空
V4 航天(宇宙航行)
[V7]航空、航天医学
X 环境科学、安全科学
X1 环境科学基础理论
X2 社会与环境
X3 环境保护管理
X4 灾害及其防治
X5 环境污染及其防治
X7 行业污染、废物处理与综合利用
X8 环境质量评价与环境监测
X9 安全科学
Z 综合性图书
Z1 丛书
Z2 百科全书、类书
Z3 词典
Z4 论文集、全集、选集、杂著
Z5 年鉴、年刊
Z6 期刊、连续性出版物
Z8 图书报刊目录、文摘、索引

2. 主题检索途径

主题检索途径是依据信息资料内容的主题属性范畴进行检索的常用途径。主题词是标引人员和检索人员的通用词。各种检索工具或检索系统所采用的全部主题词,是通过参照关系和规范化处理,使同义词、近义词、同族词、相关词作为加工与标引以及检索人员的共同依据。它打破了按学科分类的单一方法,使分散在各个学科领域里的有关课题的信息按字顺集中于同一主题范围内,使用时就如同查字典一样按字顺找到所需的主题词,在该词下,列出反映该主题内容的有关信息。主题目录和主题索引就是将文献按表征其内容特征的主题词组织起来的索引系统。利用主题途径检索时,只要根据所选用主题词的字顺(字母顺序、音序或笔画顺序等)找到所查主题词,就可查得相关信息内容。图 4-9 是“汉语主题词表”的一个实例。

主题检索途径具有适应性强、直观性强、通用性强、专指度高、检索方便等特点,不必像使用分类途径那样,先考虑需求信息或知识的所属学科范围、确定分类号等,随时可以

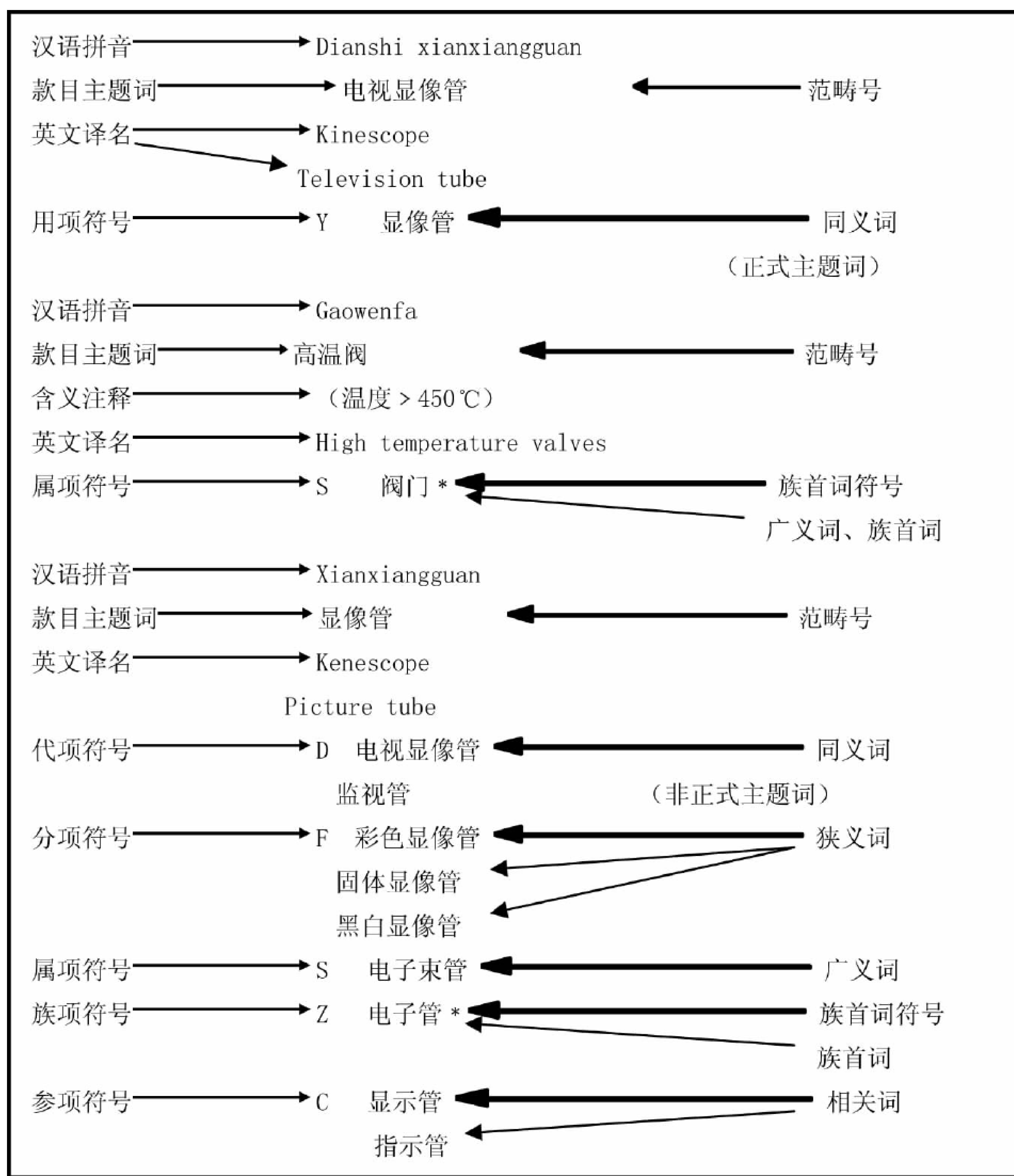


图 4-9　主题词的直观性强、通用性强与高专指度的示意图实例

增加或取消新旧信息概念主题，且具有唯一性。主题途径表征概念较为准确、灵活，不论主题多么专深都能直接表达和查找，并能满足多主题课题和交叉边缘学科检索的需要。

3. 关键词途径

该途径是按照信息标题或信息内容中具有实际意义并能表述信息主要内容、起关键作用的词或词组，按照关键词的字顺或拼音顺序在检索系统中使用的检索途径。关键词

与主题词不同之处在于：主题词是规范化的检索词，而关键词是未经过加工处理的自然语言，能够很好地表达信息生产者和信息查询的关键意图，关键词对揭示信息内容起着关键作用。

用于搜索引擎或通用数据库的信息检索，大多采用的是基于关键字索引系统（手动或者自动）组织和提取信息。其优点是无须规范化、编制索引文档快、检索入口多，缺点是由于同义词标引多，将同一主题的信息分散在不同关键词的索引文档中。

4. 题名检索途径

题名也就是信息的标题名称，例如书刊名、论文篇名、会议名称、专利名称、网页或网站标题等，用来作为检索信息途径。例如，“图书书名索引”、“期刊刊名目录”、“会议名称索引”等。题名检索的实施，需要利用题名检索工具或题名检索系统提供的题名检索功能，诸如书名目录、篇名索引、期刊名称文档等。一般多用于查找图书、期刊或单篇论文的原始文献信息。以“计算机学科的中文核心学术期刊名”为检索对象的题名检索实例如图 4-10 所示。



图 4-10 中文计算机类核心期刊题名检索实例

5. 著者途径

这是用信息的著者、编者、译者、发布者的姓名或机构团体名称作为信息的检索途径，用来检索特定的个人或团体所产生的信息。著者索引按著者姓名字顺(包括字母或笔画)编排，其信息检索直观、明了，查准率较高。国外比较重视著者途径的利用，许多检索工具和检索系统都把著者作为最基本的辅助途径。它是按著者的姓名字顺，并将有关著者生成的信息进行排序而成。以著者为线索可以系统、连续地掌握个人或机构的研究水平和信息属性动态，同一著者的信息(特别是研究性论著)往往具有一定的逻辑联系，著者途径能满足一定族性检索功能要求。

著者检索的特点是：检索者或科研人员一般都熟知自己所从事领域中的知名学者、专家、同行，以及竞争对手企业的名称，通过著者(信息生成机构名称)线索进行检索，可以系统地发现和掌握这些著者和机构的研究成果和进展的最新信息；著者或机构名称具有一定的稳定性，将其作为检索入口往往可以达到多快好省的检索效果，此外，由于著者所从事的职业、学科和专业也具有一定的稳定性，因此，还可以将著者检索看成是一种隐含的主题检索。由于著者的"同名性"(即著者姓名相同)，特别是我国同名现象普遍，在使用著者途径检索信息时，需要使用"高级检索"的分类、主题、机构、来源或职业等进行组配与逻辑功能，否则会产生大量无关信息并增加对检索结果的评估、筛选与利用的难度和工作量。

6. 序号途径

序号途径是按照信息出版或生成时所编的特征性序号来检索信息的辅助途径。这类检索有"专利号索引"、"标准号索引"、"报告号索引"等。号码一般用字母或数字或它们的混合形式来表示，检索按号码顺序查找。如美国的《化学文摘》(CA)就使用了专利号索引。利用序号途径，需对序号的编码规则和排检方法有一定的了解；往往可以从序号判断特定信息的种类、出版的年份等，有助于提高检索的查准率。

7. 分子式途径

这是以化学物质的分子式作为检索标识来检索信息的一种途径。使用的检索工具是"分子式索引"。从"分子式索引"中检索出化学物质的准确名称，然后再检索"化学物质索引"。该途径主要在美国《化学文摘》(CA)中使用。

8. 引文途径

引文途径是从作者途径去检索引用该作者著作的相关文献信息或者网络中的链接网页，它不仅反映了某个作者历年来生成了哪些信息，而且也反映了该作者的每篇信息又被哪些相关作者进行了借鉴、参考与引用，从而又进一步生成了哪些相关信息。比较常用的

检索工具有美国的《科学引文索引》(SCI)、中国社会科学引文索引等。利用引文索引可以了解某作者的某篇信息被引用的情况,进而评价某一信息的作用价值或关联性价值,以扩大信息检索范围,从而保证信息检索的查全率。

9. 特征代码途径

这是通过特征代码检索并获取特定信息的常用途径,比如大学生常常用手机扫描商品或网站的二维码就是典型的实例。特征代码包括如图书的国际标准书号(ISBN)、国际标准连续出版物代号(ISSN)、专利号、合同号或产品代码(例如商品的条形码)、读者的借阅证号、人的身份证号、网站二维码等。某些特征代码是信息类型或信息内容的特有标识,在已知信息代码的前提下,用此途径检索信息更加方便、快捷而且准确、高效。例如,利用具有全球唯一性的 ISBN 或 ISSN 可迅速地从数据库中查询特定的唯一性图书或期刊;利用 SIC 代码,可以快捷地检索出美国企业生产的产品。但代码检索的前提是需要掌握欲查询信息的代码含义,这些代码的含义和标识符往往可以利用某些检索工具或系统的辅助检索功能进行认识和把握。

10. 其他途径

除了上述常见的检索途径之外,还可按照专业领域的需要,以及文献或信息的出版类型、出版日期、出版国别、语种、所载信息的域名、IP 地址、文件路径等特征,进行信息检索。

总的来说,分类途径以学科体系为基础,按分类编排,系统性好,适合于族性检索;主题途径直接用文字表达主题,概念准确、灵活,适合于特征检索;关键词检索以自然语言的方式能够揭示信息生产者的自然意图并对表达内容含义起着关键作用。而以信息外部特征的诸多检索途径来查询信息,便于信息用户理解和识别,直观、明了、快捷且信息检索准确(例如商品条形码或二维码、网站 IP 地址等)。

4.6 信息检索方法

在浩如烟海的信息世界中要迅速、准确地查阅到自己所需要的信息,需要遵循准确、全面、深入、快捷的一般检索原则,其中首要的原则是准确。在信息检索活动中要勤于积累、善于思考。更重要的是,要灵活掌握和运用信息检索的基本方法。一般来说,信息检索的方法主要有以下几种。

1. 常规法

常规法又称检索工具法,是指直接利用检索系统(或检索工具)的方法。它是以主题、

分类、著者等途径,通过检索工具获取所需信息的一种主要方法,这种方法又可分为顺查法、倒查法和抽查法。

(1) 顺查法。顺查法是指在约束的起始年代范围内按照时间顺序,由远及近地利用检索系统逐年进行信息检索的方法。这种方法能够搜集到与某一信息需求相关的系统性内容,它适用于较大需求主题或研究课题的检索。例如,已知某需求课题的起始年代,需要掌握其发展的脉络与全过程,就可以用顺查法从课题最初研究的年代开始,逐渐向近期查找。又如,已知某项创造发明或研究成果最初产生的年代,需要了解它的演变与最新发展情况,即可从最初年代开始,按时间的先后顺序,逐年地往近期查找。用这种方法所查得的信息较为系统全面,基本上可以反映某学科专业或某课题发展的全貌。一般在申请专利的查新、准备论文开题报告、学术论文的研究综述撰写、课题论证等活动中多采用这种方法。

(2)倒查法。倒查法是由近及远,从新到旧,依据逆时间顺序检索所需信息,它的重点是放在相关需求课题的最新内容上。使用这种方法可以较快地获得最新资料,这种方法有利于保证所获得信息的新颖性,可以提高检索的效率。倒查法可以依据论文或论著的参考文献信息或者网页的相关链接页面等提示,对所需信息进行一定时间节点的追溯,对进一步启发自身的真正信息需求点,明确自身的最终信息获取目的与创新性应用,有很好的帮助作用。

(3) 抽查法。抽查法是指针对信息需求内容的某些时间段或针对需求项目的某些主题范围,选择有代表性的可能样本进行抽样检索的方法。这种方法针对性强,节省时间,信息筛选量较少、信息评价与利用效率加快,有利于提高信息检索活动的效率。抽查法的核心是样本量和抽查概率评估方法,如果抽查法的样本量不足或抽样概率简单化,则获取的信息样本所具有的代表性就不强,有可能误导真实的信息需求意图或信息利用价值。反之,抽查法的信息样本量越大或抽样概率方法越复杂,则信息检索活动的工作量和成本就会成倍增加,其信息检索结果的代表性就会大幅度提高。

对于大学生而言,顺查法、倒查法和抽查法各有优点,顺查法在时间上由远及近,查全率较高;倒查法在时间上由近及远,查准率较高;抽查法则用于满足信息需求的高级阶段(例如考研冲刺、课程期末复习、研究项目结题、学术论文总结等),信息检索的效率较高。

2. 引文法

引文法常常称为引文索引法。"引文索引法"(citation index),最初是指一种以文献之间的引证关系为基础编制的、供人们从被引证文献的角度去检索引证文献的方法,又称"引证索引"。目前已经延伸到各种数据库内部信息之间的耦合度(关联性程度)或 Web

信息之间链接层次关系(网页重要性评判)等诸多领域。引文索引法是指利用引文索引,如科学引文索引(SCI)、社会科学引文索引(SSCI)、中文社会科学引文索引(CSSCI)等,从被引论文开始查找引用它的全部论文的情况。通过这种方法可以由远及近地得到与同一主题相关的批量信息之间的关联度和彼此的重要性程度,可以使信息检索的结果“越查越新”、“越查越重要”,从而在保证信息检索查全率的基础上,获得最新、最有价值的信息。

3. 综合法

综合法又称分段法、循环法或交替法,是交替使用常规法和引文法来进行融合性检索活动的方法,它可以对常规法和引文法进行取长补短,相互配合,以获得更好的检索结果,最大化满足信息用户的信息需求。在进行具体检索时,首先利用检索工具查找出一批相关信息,再利用这些信息资料所附录的参考信息或网页链接信息进行进一步追溯查找。如此交替、循环使用常规法和引文法,不断地进行扩展查询,直到满足检索要求为止。这种方法兼有常规法和引文法的优点,使得信息的查全率和查准率都得到大幅提高。

4. 浏览法

利用以上的常规法、引文法或综合法检索信息是大学生和科研人员获得信息的主要检索途径,只要方法得当,往往可以事半功倍。但是,由于一般检索系统或检索工具只能存储有限范围的新闻、期刊图书或信息用户的自媒体信息(例如社交网络上的用户自己生成的信息),而且由于信息和知识的版权问题,检索工具与原始信息之间往往会有一定时间差。为了弥补这些缺陷,信息用户还可以借助浏览法等其他方法来收集所需要的信息。浏览法是高层次人才(包括大学生、科技工作者在内)获取信息的一种重要方法,即信息用户对本专业或本学科的重要期刊、学科网站和专门数据库等,尤其是权威核心期刊、专著和学科网站的信息进行逐一浏览查阅,以掌握最新动态和发展动向。浏览法的优点是能够及时地查阅最新生产的原始信息内容,最快地获取第一手资料,例如有规律地浏览专业或行业网站及其专题数据库,可以逐步积淀专业性信息检索的信息量基础,为日益增长的常规法、引文法或综合法信息检索提供支持作用。

4.7 信息检索策略

信息检索是一项实践性很强的活动,它要求信息用户善于思考,并通过经常性的实践,逐步掌握检索规律,从而在海量信息源中准确而高效率地检索、获取与利用信息,实现信息需求满足的最大化。

所谓信息检索策略(如图 4-11 所示),是指用户在信息需求分析的基础上拟定恰当的

检索方案,为检索过程提供潜在或快速的指导,其目的是为了优化检索过程,提高检索效率,全面、准确、快速、低成本地检索到所需信息。信息检索策略一般包括信息需求分析、选择相关信息资源、构造检索表达式、选择检索方法进行操作、对检索结果评价和对检索策略进行调整等过程。

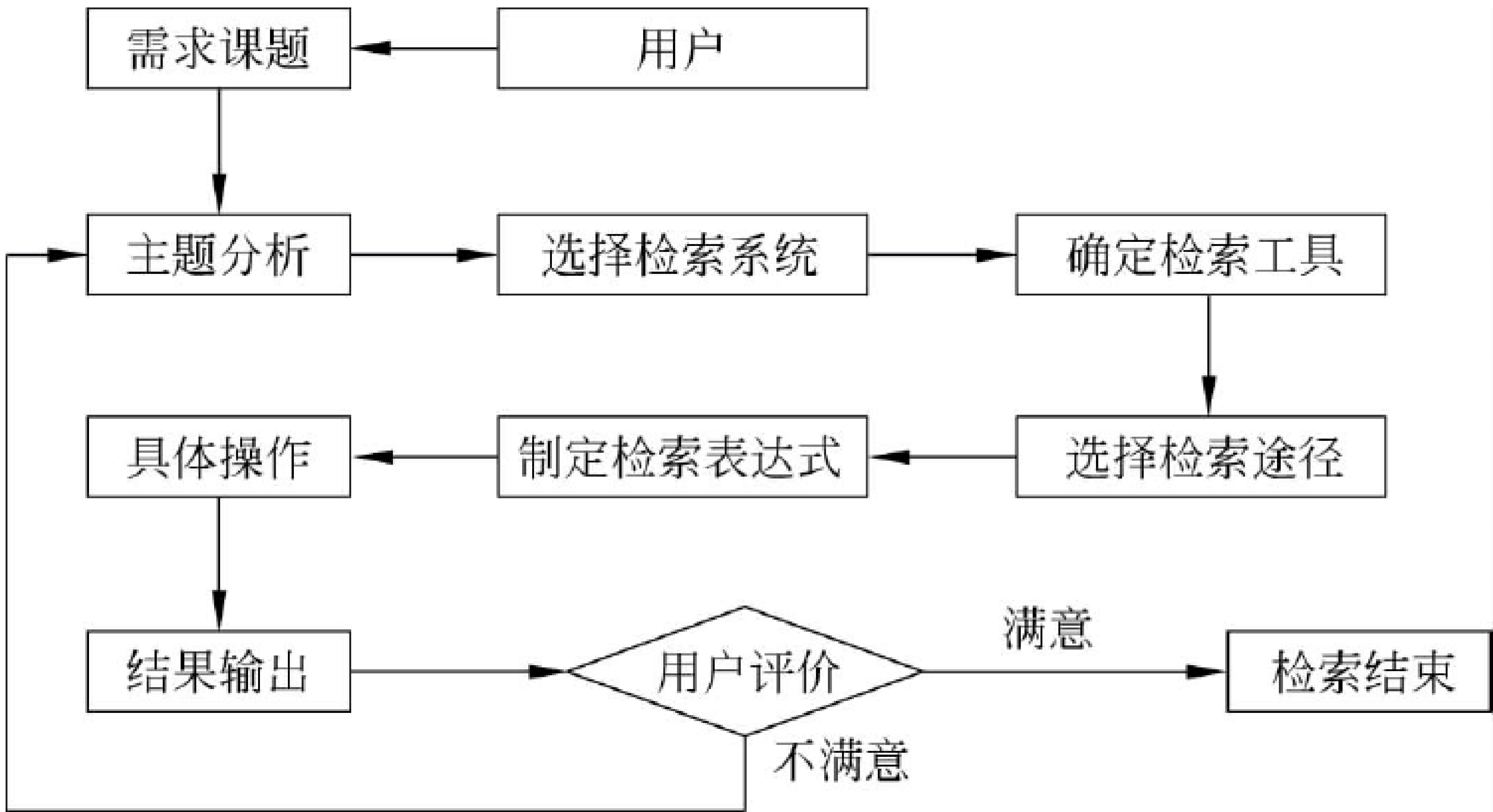


图 4-11　信息检索策略示意图

1. 信息需求分析

在信息化与网络化环境中,信息量之所以呈几何级数增长,信息需求是真正的动因。但是对信息用户个体而言,其明确的信息需求分析,不仅是信息检索过程中最首要的环节,也是高效率开展检索活动且成功满足信息需求的前提。它包括分析信息需求的主题内容、所涉及的学科范围、所需信息的资源类型、涵盖的具体时间段、检索的成本开销、可用信息资源范围、检索技术手段的可行性与利用信息的目的与要求等内容。尽管在信息化的今天,人们对信息检索策略的反应迅速而果断,但是信息需求分析的前提性作用及其需求分析的逻辑性要求是不可缺少的。

2. 选择检索工具或检索系统

在现代信息检索过程中,正确地选择检索工具或检索系统对顺利完成检索任务、保证检索质量是至关重要的。在选择检索的资源对象时,应注意选择资源的学科和专业范围,数据库存储的资源类型,信息时效年限与更新周期,数据库描述信息内容的质量、检索入口和检索语种等内容。

检索工具的种类繁多,其信息类型、学科和专业的收录范围各有侧重,所以应根据需求课题的检索要求,尽可能准确而全面地把握检索工具。检索工具的选择通常有两种方法:第一,专业性检索工具选择,例如直接选择“博士学位论文数据库”或“专利数据库”

等;第二,通用型或集成性检索工具选择,例如图书馆跨库集成检索系统或综合性网络搜索引擎的应用。

3. 选择检索方法和检索入口

一般来说,数据库等信息检索工具都提供了多种检索方法或辅助索引工具,例如顺查法、倒查法和抽查法等检索方法(检索工具大多依据信息的生成时间可供用户分区或提供辅助索引查询功能);同时,也提供了多个检索入口,例如初级检索、高级检索、专业检索等多个检索入口与相应的用户操作界面,包括许多搜索引擎也提供了简单检索、高级检索、特殊检索等检索入口。

对于大多数学生而言,应用“初级检索”去查询信息的情况较普遍,而“高级检索”或“专业检索”的应用较少,这表明学生信息检索技能缺乏,方法应用不当,信息检索效率与质量不高。

初级检索也称“傻瓜式检索”,是针对各种层次的全部信息用户而言的(例如搜索引擎),所以各种检索工具的初始检索界面一律都是“傻瓜式检索”界面。初级检索或简单检索易学易用、简单明确、界面清晰,但其检索速度最慢、信息查准率最低,因此信息用户筛选和评价检索结果的工作量大,大量的不相关信息干扰甚至误导用户的信息评价与选择。

高级检索或专业检索一般会给出较多的检索项供用户拟定能够准确反映信息需求的“逻辑检索表达式”。高级检索有时也称为逻辑组配式检索,有利于信息用户综合应用各种检索运算符或操作命令精确地构造和表达信息需求。图 4-12 是某高校图书检索系统的高级检索界面实例,但该高级检索界面的应用统计占比为 1.13%。

4. 确定检索途径

在信息需求分析的基础上,选择好信息检索工具并确定检索入口后,需要进一步明确检索途径。常用的检索途径有表达信息内容特征的分类、主题、关键词途径等,也有表达信息外表特征的题名、著者、机构、时间等途径。应注意将多种信息检索途径进行组配或逻辑组合使用,以达到更好的高级检索效果和提高信息查准率目的。

5. 检索策略调整

“检索”功能或“搜索”功能操作执行后,如果对检索结果不太满意,应及时调整检索策略。调整检索策略时,常常要利用数据库的检索限制条件、模糊/精确匹配检索、二次检索等功能,提高查准率和查全率,直到满意为止。

如果第一次检索出来的结果信息不充分,需要扩大检索范围,这时调整检索策略的方法有以下几种:①减少逻辑与算符(and),增加同义词或同族相关词,用逻辑或算符(or)将它们连接起来;②在主题词或关键词相同的词后使用截词符“?”进行扩展;③去除已有

图 4-12　某高校图书检索系统的高级检索界面

的字段限制、位置算符限制(或改用限制程度较小的位置算符)、时间约束、信息源约束(例如由限定图书扩展到期刊设置网页)等。

如果检索出来的结果信息太多,干扰了信息筛选,可以考虑增加限制条件来缩小检索范围,这时调整检索策略的方法有以下几种:①减少同义词或同族相关词;②增加限制概念,用逻辑与(and)将它们连接起来;③使用字段限制,或者限制检索词在指定的基本字段出现,或者指定辅助字段,限制结果的文献类型、语种或出版国限定等;④使用适当的位置算符;⑤使用逻辑非(not)算符,排除无关概念;⑥在结果中进行“二次检索”;⑦构造较为复杂的检索逻辑表达式。

4.8　信息检索质量与评价

信息检索质量与评价是指检索系统或检索工具实施信息检索的有效程度,它反映了用户的信息检索技能、检索工具与检索系统的性能及其信息服务质量。前文中多次出现的查全率和查准率概念,就是信息检索质量与评价的重要指标。

4.8.1 信息检索质量与评价指标

反映信息检索质量的因素主要有查全率(recall ratio)、查准率(pertinence ratio)、漏检率(omission ratio)、误检率(noise ratio)以及新颖率、检索速度等。

1. 查全率

查全率又称检全率或命中率，是指检索出的相关信息量与检索系统或检索工具中的相关信息总量的百分比。它是衡量信息检索系统或检索工具检索出相关信息能力的重要尺度，定义为

$$查全率=\frac{检出的相关信息量}{系统中相关信息总量}\times 100\%$$

2. 查准率

查准率又称检准率或相关率，是指检索出的相关信息量与检索出的信息总量的百分比。它是衡量信息检索系统或检索工具的精确度的指标，可以定义为

$$查准率=\frac{检出的相关信息量}{检出的信息总量}\times 100\%$$

3. 漏检率

漏检率又称漏检概率，是指未检索出的相关信息量与检索系统中相关信息总量的百分比。它是与查全率相对应的概念，即“漏检率＝100％－查全率”。漏检率是衡量信息检索系统漏检信息的尺度，可以定义为

$$漏检率=\frac{未检出的相关信息量}{系统中相关信息总量}\times 100\%$$

4. 误检率

误检率又称检索噪声，是指检索出结果中不相关信息量占检索出信息量的百分比。它是与查准率相对应的概念，即“误检率＝100％－查准率”。误检率是衡量信息检索系统误检信息的程度的指标，可以定义为

$$误检率=\frac{检出的不相关信息量}{检出的信息总量}\times 100\%$$

根据有关实验表明，查全率与查准率是成反比关系的，是相互制约的。一般认为，一个检索系统或检索工具的查全率在60％～70％，查准率在40％～50％即能满足用户信息需要，100％只是理论上的标准，在实际检索活动与检索系统中不可能达到这一理想状态。

5. 新颖率与检索速度

新颖率指的是在检索结果中最新信息所占的比重。其中“最新信息”一般指所需信息

中最近一段时间(例如最近一个月、最近一个季度或最近一年)的信息。

$$新颖率=\frac{最新信息量}{检出的信息总量}\times 100\%$$

检索速度也称为检索反应时间或检索响应时间,它是用户拟定信息需求主题、确定检索方法与途径、选择检索工具或检索系统对象、检索工具或系统的数据处理和网络反馈传输、用户筛选和提取所需信息等过程的时间总和。检索速度与信息用户的检索素养和检索系统的处理与反馈性能密切相关。

4.8.2 影响检索效果的因素

查全率与查准率是评价检索效果的两项重要指标。它们与信息资源的存储与检索两个方面直接相关,即与系统的信息采集范围、标识规范、标引工作和检索工作等都有着非常密切的关系。

1. 影响查全率的因素

从信息存储的角度,影响查全率的因素有以下几个。

(1) 影响查全率的因素主要有检索系统采集信息的范围有限。

(2) 检索系统不具备截词和信息自反馈功能或自适应能力较低,建立索引的方法不够完善。

(3) 主题词或关键词结构体系不完整,词汇缺乏深入控制和专指性,词间关系模糊或不正确。

(4) 信息的人工标识或自动标识质量不高,出现标识前后不一致或标识工作人员遗漏了重要概念或用词不当等情况。

从信息检索过程来看,其影响因素主要有以下几个。

(1) 信息用户的信息检索素养不高。

(2) 检索策略过于简单。

(3) 选择主题词或关键词及其逻辑组配不当。

(4) 使用的检索途径和方法太少。

(5) 依赖初级检索过多,不能全面地描述检索需求等。

2. 影响查准率的因素

影响查准率的因素主要有以下几个。

(1) 检索式中“逻辑或”的使用或者“截词”部位(包括前截词、中间截词和后截词)不当。

(2) 检索系统或检索工具不具备逻辑非的功能、二次检索功能。

(3) 索引词不能准确地描述信息主题词和关键词及它们之间的良好层次关系。

(4) 检索时所用检索词(或检索式)的专指度不强,检索面过宽,选词及词间关系不正确。

(5) 信息标识过于详尽,组配规则不严密或出现逻辑错误等。

实际上,影响检索效果的因素是非常复杂的。要想达到较高的查全率,势必需要对检索范围和一些限制条件逐步放宽,其结果是会把很多不相关的信息也带入数据库系统,影响了查准率。要想同时提高查全率和查准率是不容易的,而强调一方面、忽视另一方面也是不妥当的。信息用户应当根据具体的信息需求,合理地调节查全率和查准率,以保证获得更好的检索质量。

3. 提高信息检索质量的措施

提高查全率的措施通常有以下几种。

(1) 使用泛指度较强的检索主题词或关键词(如上位簇首词或上位主题词)。

(2) 将待检索的信息需求中同一概念面的同义词、近义词及相关概念充分列举,并用布尔运算符逻辑或进行组配。

(3) 使用截词符“?”或“*”。

(4) 改变检索项,例如当要求检索词位于标题中或为关键词或主题词时,检出的记录数就会比较少,这时可改为要求检索词位于摘要或全文中,检索出的信息数量即可增加。

(5) 减少限制条件,增加近似检索项。

提高查准率的措施通常有以下几种。

(1) 使用专指性较强的检索主题词或关键词(如下位类或下位主题词)。

(2) 增加检索词之间的互相限定,并用布尔运算符“逻辑与”进行组配。

(3) 少使用截词符“*”或“?”。

(4) 改变检索项。例如,当要求检索词位于摘要或全文中时,检索出信息数量较多,则可改为要求检索词位于标题中。

(5) 缩减信息的时间范围、语种、国别或信息源属性(期刊、论文、标准等)范围等的限制条件。

(6) 选择专业性高且权威性强的检索系统与检索工具,例如专利数据库、优秀博士学位论文数据库、美国的 SCI 等。

本章小结

信息源也就是我们在检索过程中经常接触到的不同信息集合或者不同检索对象实体，也就是我们获得原始信息内容的来源。依据信息内容的加工层次划分，信息源范围包括零次信息源、一次信息源、二次信息源、三次信息源。信息源的出版发行与共享类型主要有图书、期刊、会议文献、科技报告、专利文献、学位论文、技术标准、政府出版物、产品样品和说明书、技术档案、报纸等，要注意这些外文信息源的准确识别与利用。

信息检索工具是以压缩形式存储、报道和查找信息线索或原始信息全文的工具，它是经过对信息进行搜索整理、特征分析和组织加工后的产物，同时又是信息检索的主要手段和条件。它包括传统的检索工具，例如科学引文索引 SCI；也包括网络检索工具，例如 Baidu 等。信息检索工具的主要功能表现在存储和检索两个方面。信息报道及时、全面，存储规范、有序以及检索迅速和准确是对检索工具的基本要求。

信息检索工具主要有目录、索引、文摘、参考工具书和搜索引擎。信息检索途径多种多样，其中表明信息外部特征的相关途径有标题、责任者(或作者与发布者)、产生机构、序号、信息来源、产生时间、范围、路径、点击量或访问量等途径；与信息内容特征相关的途径有学科分类、主题和关键词、内容代码(例如化学分子式、图像色彩等)等途径。信息检索方法主要有顺查法、倒查法、抽查法、引文法、综合法、浏览法等。

信息检索策略，是指用户在信息需求分析的基础上拟定恰当的检索方案，为检索过程提供潜在或快速的指导，其目的是为了优化检索过程，提高检索效率，全面、准确、快速、低成本地检索到所需信息。信息检索策略一般包括信息需求分析、选择相关信息资源、构造检索表达式、选择检索方法进行操作、对检索结果评价和对检索策略进行调整等过程。

信息检索质量与评价是指检索系统或检索工具实施信息检索的有效程度，它反映了用户的信息检索技能、检索工具与检索系统的性能及其信息服务质量，其中查准率和查全率是主要的评价指标。

本章思考与练习题

1. 信息内容的加工层次划分有哪些信息源？分别举例说明。

2. 查询并举例下列信息源：图书、期刊、会议文献、科技报告、专利文献、学位论文、技术标准。

3. 分别举例说明英文信息源如何辨别。
4. 用图示说明检索工具的含义。
5. 常用的检索工具有哪些？分别举例。
6. 信息检索有哪些主要途径？分别举例。
7. 信息检索有哪些主要方法？
8. 用图示说明信息检索一般策略。
9. 信息检索质量有哪些评价指标？请举例说明。
10. 良好的信息检索方法与策略对信息检索素养的形成有何作用？

第二部分

信息检索素养基本原理篇

本部分包括6章内容(第5章至第10章),从信息检索的基本原理即信息检索的主要内在工作机制与技术方法以及依据的数学逻辑知识等内容,来进一步培养大学生的信息检索素养。大学生在信息检索与利用过程中,无论是基于普通的"初级检索"(基本检索、一般检索或通用检索等),还是基于较高检索需求的"高级检索"(复合检索、主题式检索或复杂逻辑检索等),甚至是针对性强和专业性要求高的"专业检索"(专门检索或专家检索等),都需要学习和掌握信息检索基础数学原理、文本分类与文本索引构建、图像信息检索、音频信息检索、视频信息检索和Web信息搜索等基础理论知识。第二部分的学习与掌握,不仅是当代大学生(尤其是研究生)信息检索素养教育的重要内容,也是大学生与其他社会群体的信息检索素养相互区别的重要内容。

第5章"信息检索的基础数学原理"的引入,使得信息检索有了更加严谨的逻辑论证,检索过程和信息需求的本质描述也更为精确,从而使得信息检索的理论与实践获得持续性的基础支撑。内容包括布尔检索、检索的检索模糊集合论、扩展布尔检索、信息检索向量空间模型、潜在语义索引模型、神经网络检索模型、概率论检索模型、检索粗糙集理论、检索遗传算法等。

第6章论述了文本分类与文本索引构建。文本分类(text categorization, TC)又称为文本自动分类,它是信息检索和文本数据挖掘的重要基础。文本

自动分类能较好地解决大量文档信息归类问题并应用到很多信息领域。文档是建立各种文本型检索数据库的基础，从组织形式上划分，文档可以分为顺排文档(sequential file)和倒排文档(inverted file)两种。倒排文档就是把顺排文档中具有检索属性的项目信息抽取出来，重新排列组织成新的数据文档，在很多数据库中被称为索引文档。

第7章、第8章和第9章分别阐述了图像信息检索、音频信息检索和视频信息检索的基础性原理。随着因特网和移动互联网的快速发展，数据量庞大的图像、音频和视频等多媒体信息资源日益成为网络用户的重要查询与利用对象。与传统基于文本的信息检索原理不同，图像、音频和视频等多媒体信息资源主要是基于内容的信息检索，而且基于内容的多媒体信息检索的检索精度也要高得多，因此备受重视而成为大学生信息检索素养教育不可或缺的主要内容。

第10章论述了Web信息搜索的一般性原理。Web是WWW(万维网)的简称，它是Internet最基本、最广泛的应用服务，也是最主要的信息资源类型。对于信息社会和网络时代的信息用户而言，直接面对的Web信息获取工具就是网络搜索引擎，Google、Baidu等搜索引擎是Web信息采集与搜索的典型代表。

第5章　信息检索的基础数学原理

由于当今信息量呈几何级数膨胀和用户信息需求多样化发展趋势，在检索的实践活动中会涉及大量的信息处理与存储过程。用户信息检索的最终实现必须依靠强有力的计算机应用程序去自动执行或智能信息处理作为支撑，而强有力的计算机应用程序必须依据数学原理及其模型方法的建立为前提，利用数学原理与模型方法来建立检索基础模型是必不可少的工作。运用数学原理不仅能使信息检索作为研究对象的概念含义精确化，而且能够深刻揭示信息检索过程的显性现象与潜在的隐性规律。在信息检索中引入数学原理及其模型方法，将检索过程中的信息及其处理过程加以解释和抽象，表达成某种数学模型，再经演绎与推断，从而指导检索实践和促进检索工作的技术进步。数学原理及其模型的引入使得信息检索有了更加严谨的论证，检索过程和信息需求本质的描述也更为精确。迄今为止，基于集合理论的布尔模型、Salton 模型和模糊集合模型等数学一般原理最为成熟，也在检索实践中得到了普遍应用。

5.1　简单布尔检索

5.1.1　基本原理

布尔模型是一种以经典集合论和布尔代数为理论基础的非常简单的信息检索模型。它采用布尔代数的方法，用布尔逻辑表达式表示用户需求提问，通过对信息标识和提问式的比较来检索信息。对某一特定的信息，通常表示成 $D=(t_1,t_2,\cdots,t_n)$ 的形式。由于布尔逻辑式可以表达成与用户思维习惯相一致的提问要求，因此，用户提问可以表示为由三种逻辑运算符即逻辑与(*)、逻辑或(+)和逻辑非(－)连接起来的布尔表达式，标引词 t_1 和 t_2 之间可能具有的逻辑运算是 $t_1\wedge t_2$ 和 $t_1\vee t_2$，而任一标引词的逻辑非运算为 $-t$，这些逻辑运算将作为用户提问的一部分出现在布尔表达式的某个位置上，图 5-1 可以很直观地显示这些逻辑运算的结果。

显然，上述的布尔运算实际上是集合之间的交、并、补运算。也就是说，布尔检索实际上是通过若干个检索词所包含的信息集合的交、并、补运算来响应用户信息需求提问的。

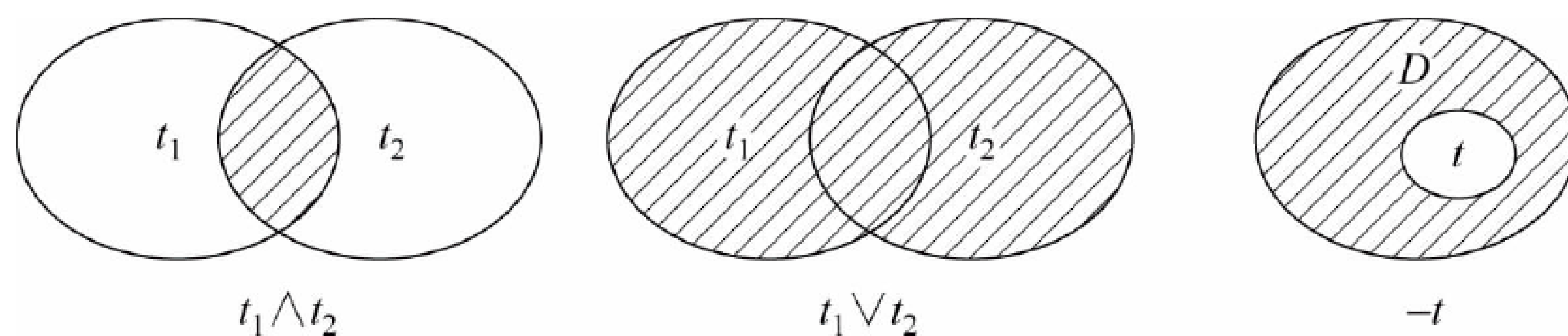


图 5-1　布尔运算逻辑关系图

布尔模型在解释信息检索的数据处理过程时，主要遵循两条基本规则。

系统索引词集合中的每一个索引词在一篇文档中只有两种状态：出现或者不出现。相应地，每个索引词的权值 $w_{ij} \in \{0,1\}$。

检索提问式 $\boldsymbol{q}$ 由三种布尔逻辑运算符"and"、"or"、"not"连接索引词来构成。

根据布尔逻辑的运算规定，提问式 $\boldsymbol{q}$ 可以被表示成由合取子项（conjunctive components）组成的析取范式(disjunctive normal form，dnf 或 DNF)形式。例如，布尔提问式

$$\boldsymbol{q}=k_1 \text{ and}(k_2 \text{ or not } k_3)$$

可以写成如下等价的析取范式形式：

$$\boldsymbol{q}_{\text{dnf}}=(k_1 \text{ and } k_2 \text{ and } k_3)\text{or}(k_1 \text{ and } k_2 \text{ and not } k_3)\text{or}(k_1 \text{ and not } k_2 \text{ and not } k_3)$$

这里，$\boldsymbol{q}_{\text{dnf}}$ 为提问式 q 的主析取范式。进一步地，可以用如下简化形式来表示 $\boldsymbol{q}_{\text{dnf}}$：

$$\boldsymbol{q}_{\text{dnf}}=(1,1,1) \text{ or } (1,1,0) \text{ or } (1,0,0)$$

其中，(1,1,1)、(1,1,0)和(1,0,0)是 $\boldsymbol{q}_{\text{dnf}}$ 的三个合取子项（合取子项可用符号 $\boldsymbol{q}_{\text{cc}}$ 表示），它们是一组向量，由对应三元组(k_1,k_2,k_3)的每一分量取 0 或 1 值而得到。

基于上述规则与假定，布尔模型对于任一篇文档 $\boldsymbol{d}_j \in D$，定义 $\boldsymbol{d}_j$ 与用户提问 $\boldsymbol{q}$ 的匹配函数为

$$\text{sim}(\boldsymbol{d}_i,\boldsymbol{q}) = \begin{cases} 1\text{，如果存在 } \boldsymbol{q}_{\text{cc}} \mid (\boldsymbol{q}_{\text{cc}} \in \boldsymbol{q}_{\text{dnf}}) \text{ 且对于任意 } k_i\text{，有 } g_i(\boldsymbol{d}_j) = g_i(\boldsymbol{q}_{\text{cc}}) \\ 1\text{，其他} \end{cases} \tag{5-1}$$

式(5-1)中，函数 g_i 定义为 $g_i(\boldsymbol{d}_j)=w_{ij}$。现在，假设文档集合 D 中存在两篇文档 $\boldsymbol{d}_1$ 和 $\boldsymbol{d}_2$，其中，$\boldsymbol{d}_1$ 含有索引词 k_1 和 k_2，$\boldsymbol{d}_2$ 含有索引词 k_1 和 k_3，则它们的文档向量分别为

$$\boldsymbol{d}_1 = (1,1,0)$$

$$\boldsymbol{d}_2 = (1,0,1)$$

根据匹配函数 $\text{sim}(\boldsymbol{d}_j,\boldsymbol{q})$的定义，很显然文档 $\boldsymbol{d}_1$ 与提问式 $\boldsymbol{q}=k_1\text{and}(k_2 \text{ or not } k_3)$的匹配函数值为 1，即文档 $\boldsymbol{d}_1$ 与提问 $\boldsymbol{q}$ 是相关的；而文档 $\boldsymbol{d}_2$ 与提问 $\boldsymbol{q}$ 的匹配函数值为 0，表

明文档 d_2 与提问 q 是不相关的。

5.1.2　布尔检索模型的特点

布尔模型是最早提出的一种信息检索一般数学模型。1957 年，巴・希列尔（Y. Bar-Hille）就对布尔逻辑应用于计算机信息检索的可能性进行了探讨；20 世纪 60 年代末期，布尔检索模型正式被大型文献检索系统所采用；70 年代时逐渐成为各种商业性联机检索服务系统的标准检索模式。目前，基于布尔检索框架的各类检索系统仍具有顽强的生命力，并在信息搜索与信息服务领域占据重要地位。

在布尔检索中，用户的查询要求用普通的语言叙述，即用户可完全按照自己的思维习惯提问。其中查询要求（条件）A、B、C、D 等可以分别用若干个标引词来表示，然后可以用布尔逻辑运算符“∨”、“∧”、“—”将用户的提问“解析”成信息服务系统可以接受的形式。这种结构化的提问方式与用户的思维习惯相一致，所以成为布尔逻辑检索的一个突出优点。布尔检索的一个用户界面实例如图 5-2 所示。

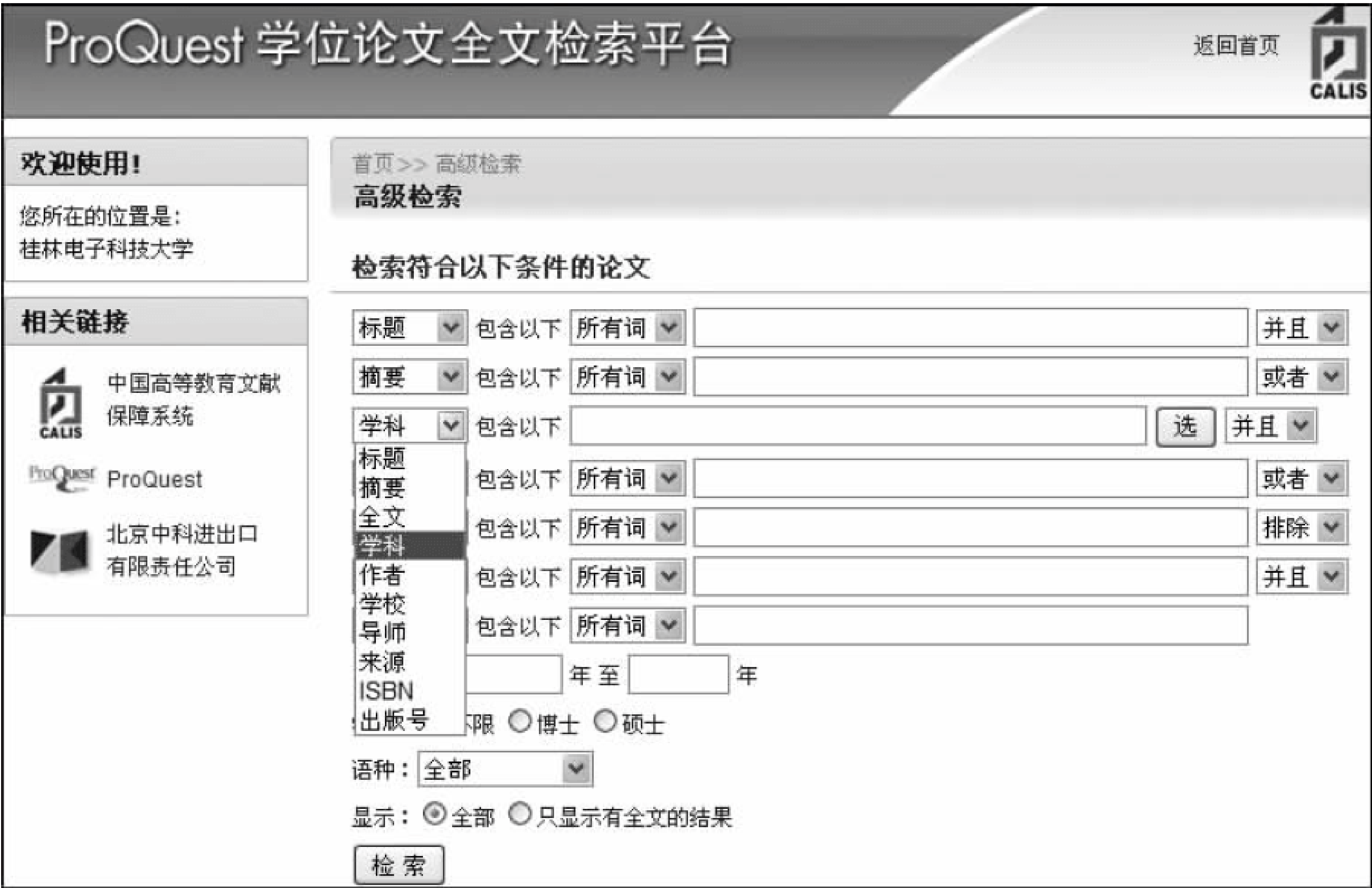


图 5-2　布尔检索实例图（以 ProQuest 为例）

以 ProQuest 为例，图 5-2 布尔检索实例图中的“并且”、“或者”与“排除”运算，就是典型的布尔检索应用。这种模型把复杂的检索过程简单化，能够将比较复杂的信息提问按其概念组配的逻辑关系描述出来，从而变成可以由计算机执行的逻辑运算，变成机器根据

事先确定的程序进行自动匹配的过程,这种运算上的简单易行是布尔逻辑检索系统的突出优势。

布尔模型具有简单性(simplicity)、容易理解性(easy understanding)、简洁形式化(clean formalism)等突出优点。布尔模型的简单性、易理解性与易实现等特点为其在检索系统和检索工具中的广泛应用奠定了良好基础。尽管布尔模型有着种种优点,但它还是存在明显的局限性。

(1) 布尔模型是基于二值判定为标准的,信息对象要么相关,要么不相关,并没有一个相关信息级别的概念,例如符合信息需要的相关性程度大小,因此很难有好的检索效果。

(2) 构造布尔逻辑式不是一件轻松的事情,对于普通信息用户,很难用 AND(逻辑与)、OR(逻辑或)、NOT(逻辑非)运算的结合来准确地表达自己的信息需求,并且检索词的简单组配也不能完全反映实际需要。

(3) 检索结果输出完全依赖于布尔提问与检索系统中信息的匹配情况,很难控制输出量的大小。

(4) 布尔提问表示存在某些不合理的地方。对于"∨"提问,包含一个在提问中出现的检索词的信息与包含几个在提问中出现的标引词的信息被认为是一样的重要;对于"∧"提问,包含多个标引词的信息与不包含任何标引词的信息被看成是一样不相关。

(5) 检索结果不能按用户定义的重要性排序输出,用户只能从头到尾浏览输出结果才能知道哪些信息更适合自己的需要。

鉴于布尔模型的这些不足,人们提出用语词加权和部分匹配的功能来扩展经典的布尔模型,将向量模型和布尔模型融为一体,来克服传统布尔模型的一些缺陷,这就是扩展布尔模型。

5.2 信息检索模糊集合论

信息检索模糊集合模型是建立在模糊集合论基础上的,模糊集合论可以看做是经典集合论的推广。1965 年美国加州大学伯克利分校的札德(LA. Zadeh)教授发表了一篇关于"模糊集合"的著名论文,由此奠定了模糊理论的研究与发展。

模糊集合论对经典集合论的推广主要表现在:它把元素属于集合的概念模糊化,承认集合论范围内存在既不完全属于某集合,又不完全不属于某集合的元素,即变经典集合论"绝对的属于"概念为"相对的属于"概念;同时,又进一步把属于概念数量化,承认论域

上的不同元素对于同一集合具有不同的隶属程度，因此引入了隶属度（membership）的概念。

模糊集合理论处理的是边界不明确的集合表示，其中心思想是把集合中的元素和隶属函数结合在一起。隶属函数的取值在[0,1]上，0 表示元素不隶属于该集合，1 表示完全隶属于该集合，值在 0 和 1 之间表示元素为该集合的边际元素。

定义：给定论域 U，U 的模糊子集 A 可以定义为 U 到闭区间[0,1]上的一个映射：LA：$U\rightarrow[0,1]$，LA 为 A 的隶属度。正如经典集合论是传统精确数学的基础一样，模糊子集论是模糊理论的基础，同样也可以定义模糊子集上的运算。常见的三种运算分别是模糊集合的补运算、两个或多个集合的并、交运算。

定义：给定论域 U，A 和 B 分别为 U 的两个模糊子集，A－是 A 关于 U 的补集，u 为 U 中的元素，则

$$\mu_{\overline{A}}(u) = 1 - \mu_{\overline{A}}(u)$$

$$\mu_{A} \cup B(u) = \max\ (\mu_{A}(u) \cdot \mu_{B}(u))$$

$$\mu_{A} \cap B(u) = \min\ (\mu_{A}(u) \cdot \mu_{B}(u))$$

5.2.1　模糊检索的数学描述

模糊检索是将信息文档看成是与提问在一定程度上相关，对于每一个标引词，都存在一个模糊的信息集合与之相关；对于某一给定的标引词，用隶属函数表示每一则信息文档与该词相关的程度，即隶属度，其取值在[0,1]上，则有信息文档 d 和标引词 t，d 对于 t 的隶属度可以定义为

$$\mu_F: D \times T \rightarrow [0,1],$$

$$(d,t) \rightarrow \mu_F(d,t)\ \forall (d,t) \in D \times T$$

则在信息检索系统中文档 d 与标引词 t 的二元模糊关系 F 可以描述为

$$F = \{[(d,t),\mu_F(d,t)]d \in D, t \in T\} \tag{5-2}$$

由于用户通常希望检索出的信息能较高地满足其需求主题，因此，这里所定义的 $\mu_F(d,t)$表示文献 d 涉及标引词 t 所达到的程度，而不是标引词 t 反映文献 d 的主题内容的程度。

标引词的模糊集合是在标引过程中建立的，标引人员不是简单地把标引词赋予信息文档，还要指出标引词与信息文档的相关程度。如 $d=\{(t_1,0.5),(t_2,0.8)\}$，数字 0.5 和 0.8 表示信息文档对于标引词 t_1，t_2 的隶属度，数值越大表示隶属度越大。当全部信息文

档标引完毕，也就为每个标引词定义了一种隶属函数，指明了每一信息文档对于每个标引词的相关程度。

隶属函数是模糊集合论乃至整个模糊学的最基本概念之一，正确构造隶属函数是应用模糊学方法的关键。由于隶属度的确定，既有客观性的一面，也有主观性的一面，因此，在解决实际问题时，构造切合实际的隶属函数至今还没有非常满意的解决方法。

5.2.2 信息文档对标引词的隶属度

在标引词集合中，由于概念相关的模糊性，两个标引词在不同程度上总是存在着语义上的关联，因此，信息文档对标引词的隶属度是通过标引词表来计算的。标引词表可以通过词-词关联矩阵来建立，这个矩阵的行和列分别对应于集合中的标引词，矩阵中词 t_i 和 t_j 的关联因子可以定义为

$$C_{i,j} = \frac{n_{i,j}}{n_i + n_j - n_{i,j}} \tag{5-3}$$

式中 n_i 表示包含标引词 t_i 的信息文档的数目，n_j 表示包含标引词 t_j 的信息文档的数目，则标引词 t 的模糊集合中，文献 d 的隶属度：

$$\mu_F = 1 - \Pi(1 - C_{i,j}) \tag{5-4}$$

5.2.3 提问检索词的相关性描述

用户提问通常是由布尔逻辑式表达的，即用布尔逻辑运算符将标引词连接起来。布尔逻辑的常用运算符有“与”、“或”、“非”，即 $\wedge$，$\vee$，$\rightarrow$。提问匹配以通过引入模糊算符来确定信息文档对于提问的相关程度。设 D 为信息文档集，Q 为提问集，$\forall d \in D, q \in Q$，$Q \times D$ 上的模糊关系 R：

$$R = \{(q,d,\mu(q,d)) \mid q \in Q, d \in D\}$$

式中 $\mu(q,d)$表示信息文档 d 对于提问 q 的相关程度。

根据模糊集合的运算规则，将三个基本的模糊运算符分别定义如下。

（1）若 $q = a \vee b$，则 $\mu(q,d) = \max(\mu(d,a), \mu(d,b))$，这里 $a, b \in T$，$\mu(d,a)$，$\mu(d,b)$分别表示信息文档 d 论述标引词 a 和 b 所达到的程度。

（2）若 $q = a \wedge b$，则 $\mu(q,d) = \min(\mu(d,a), \mu(d,b))$。

（3）若 $q = -a$，则 $\mu(q,d) = 1 - \mu(d,a)$。

在模糊集合检索中，对于布尔模型的用户信息需求的处理通常是把表达用户需求的布尔逻辑式转换成析取范式的形式。例如，$q = t_a \wedge (t_b \vee -t_c)$，可以写成与之等价的析取

范式：$q_{\text{dnf}}=(1,1,1)\vee(1,1,0)\vee(1,0,0)$，其中的每个分量都是$(t_a,t_b,t_c)$的一个二值加权向量，它们构成了$\bar{q}_{\text{dnf}}$的合取分量，用$CC_i$表示第$i$个合取分量，则提问可以推广为$p$个合取分量的形式：

$$\bar{q}_{\text{dnf}}=CC_1\vee CC_2\vee\cdots\vee CC_p \tag{5-5}$$

计算信息文档与提问相关的过程类似于经典布尔模型中的计算，只不过在模糊检索中处理的对象是模糊集合而不是普通的集合。

对于上述的提问$q=t_a\wedge(t_a\vee t_b)$，D_a表示标引词t_a在文献集上的模糊子集，它由隶属度大于既定阈值的文献所组成。同理，可以定义标引词t_a和t_c的模糊子集D_b、D_c，由于所有的集合都是模糊不确定的，即使信息文档d不包括标引词t_a，该信息文档也有可能属于集合D_a（见图 5-3）。

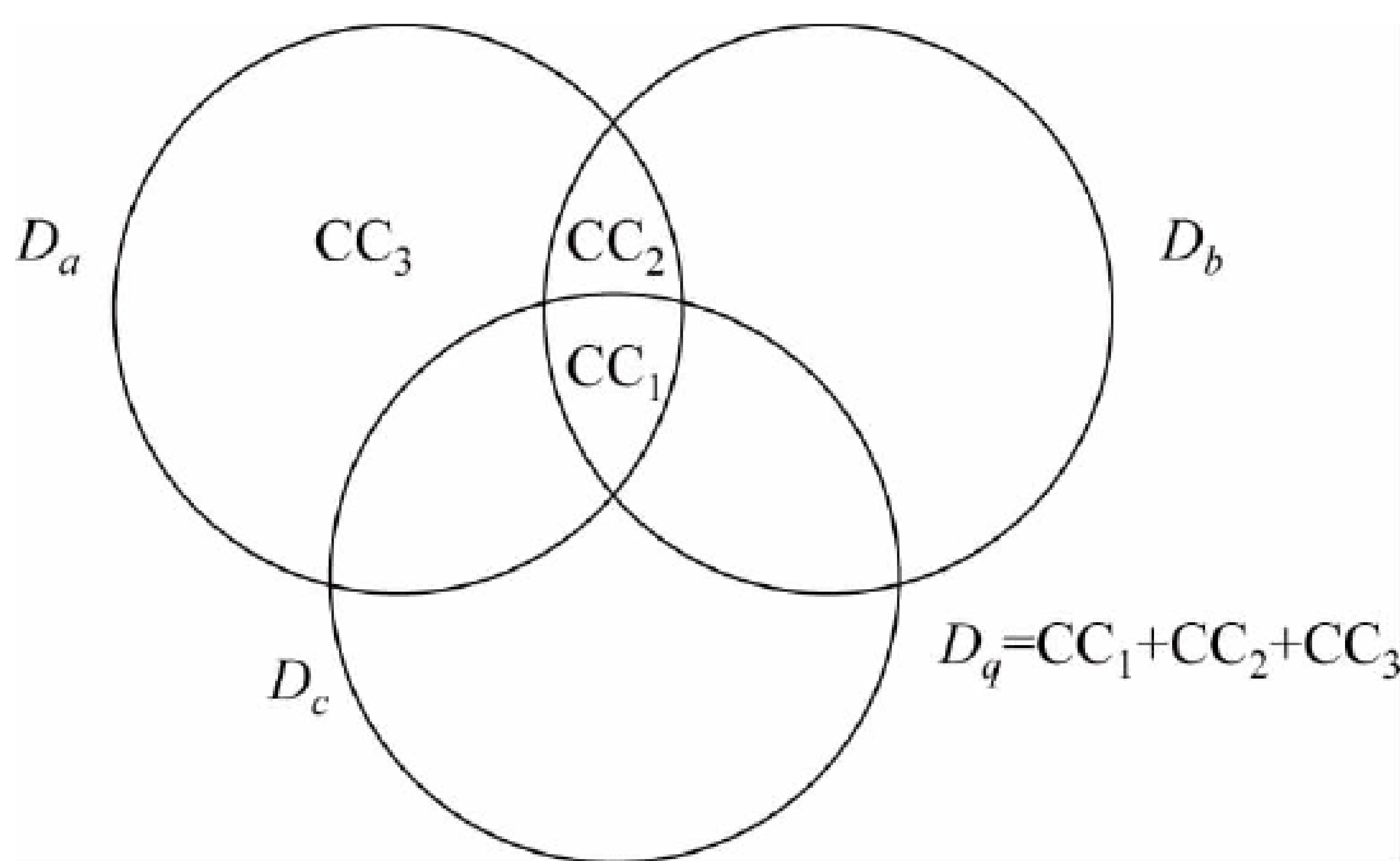


图 5-3　提问$q=t_a\wedge(t_b\vee t_c)$的模糊文献集

提问模糊集合D_q是q_{dnf}的三个合取分量的模糊集合的并运算，则D_q中信息文档d的隶属度：

$$\begin{aligned}\mu(q,d)&=\mu cc_1+cc_2+cc_3\cdot d=1-\prod_{i=1}^{3}(1-\mu cc_i\cdot d)\\&=1-\{\mu(d,a)\mu(d,b)\mu(d,c)\}\times\{1-\mu(d,a)\mu(d,b)(1-\mu(d,c))\}\\&\quad\times\{1-\mu(d,a)(1-\mu(d,b))(1-\mu(d,c))\}\end{aligned}$$

计算得出$\mu(q,d)$，它所反映的正是信息文档d对于提问q的相关程度。所以，提问q可以定义为信息文档集合D上的一个模糊子集：$q=\{(d,\mu(q.d))\mid d\in D\}$。用户给定一个阈值$\lambda(0\leqslant\lambda\leqslant1)$，将小于$\lambda$的项去掉。当$\mu(q,d)\geqslant\lambda$时，$d$作为命中的信息文档输出，输出可以采取按照对提问的相关程度的大小形式排序输出。通过控制λ的取值，可以输出合适的文献。

基于模糊集合模型的检索结果是建立在信息文档集上的，且其隶属度就是信息文档集对用户提问的相关程度的模糊子集。就目前的水平而言，还无法十分精确、有效地确定这个隶属函数：在提问匹配中引入的 max 和 min 算符不能很好地反映真实的匹配过程，而把提问的布尔逻辑表达式转换成析取范式，用代数和、代数积分计算析取模糊集合以获取模糊集合中信息文档的隶属度，更加适合于模糊信息检索应用。

模糊检索模型与经典布尔模型关系密切，它基本保留了布尔检索功能，但是更为灵活，对那些既想利用布尔检索长处，又想避免其二值相关性测度局限性的人们来说，能够较好地满足信息检索需求。模糊检索模型还支持对命中文档按相关度大小的排序输出。

5.3 扩展布尔检索

1983 年信息检索专家萨尔顿(G. Salton)及其博士生福克斯(E. A. Fox)等人提出的一种基于布尔逻辑框架的混合布尔与向量特性的混合检索模型，即扩展布尔模型。扩展的布尔检索模型是基于布尔逻辑基本假设的改进，下面采用矢量的方法来讨论布尔信息检索。

5.3.1 基于两个标引词的情形

假定信息文档集合中的信息 $\boldsymbol{d}_j$ 仅用两个标引词 t_x 和 t_y 标引，并且 t_x，t_y 允许被赋予一定的权值，其权值分别为 $W_{x,j}$、$W_{y,j}$，权值的取值范围为[0，1]，权值越接近于 1，说明该词越能反映文本的内容，反之，反映文本的内容较差。给标引词加权通常采用的是著名的 tf-idf 加权方案：

$$W_{x,j} = f_{x,j} \times \frac{\mathrm{idf}_x}{\max x_i \times \mathrm{idf}_x} \tag{5-6}$$

式中 $f_{x,j}$ 为标引词 t_x 在文献 $\boldsymbol{d}_j$ 中出现的频率，idf_x 为逆信息文档词频。为了简单起见，用 x，y 分别表示权值 $W_{x,j}$、$W_{y,j}$。我们采用二维图来表示信息文档的提问，用距离的概念表示信息文档与提问的相似度。见图 5-4。

对于析取提问 $\boldsymbol{q}=t_x \vee t_y$，只有 A、B、C 三点所代表的信息文档才是最理想的，对于任一信息文档 $\boldsymbol{D}_j$ 而言，当它离 A、B、C 三点越接近时，说明相似度越大，因而 $\boldsymbol{D}_j$ 到点(0，0)的矢量距离可以用来度量与提问 $\boldsymbol{q}_{\mathrm{or}}$ 的相似度，则

$$|\boldsymbol{D}_j| = \sqrt{x^2 + y^2} \tag{5-7}$$

显然，$0 \leqslant |\boldsymbol{D}_j| \leqslant 1$，为了使相似度控制在 0 和 1 之间，相似度可以规范化为

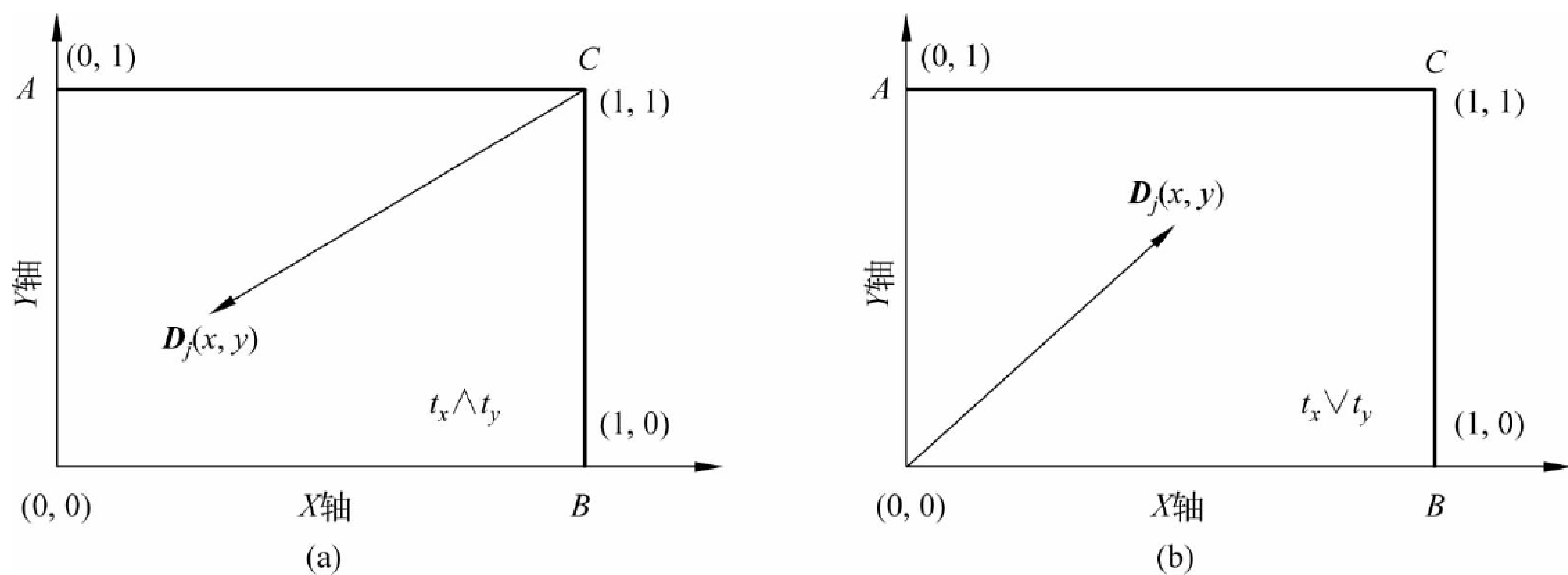


图 5-4　扩展布尔逻辑的矢量表示

$$\mathrm{sim}(\boldsymbol{q}_{\mathrm{or}},\boldsymbol{d}_j)=\sqrt{\frac{x^2+y^2}{2}} \tag{5-8}$$

对于合取提问 $\boldsymbol{q}=t_x \wedge t_y$，只有 C 点才是最理想的文献，则 $\boldsymbol{D}_j$ 到 C 点的矢量距离为

$$|\boldsymbol{D}_j|=\sqrt{(1-x^2)+(1-y^2)} \tag{5-9}$$

它可以作为衡量文献与提问之间相似度的一个尺度，则相似度可以规范化为

$$\mathrm{sim}(\boldsymbol{q}_{\mathrm{or}},\boldsymbol{d}_j)=1-\sqrt{\frac{(1-x^2)+(1-y^2)}{2}} \tag{5-10}$$

5.3.2　推广到 n 个标引词空间

以上讨论的是两个标引词的情况，信息文档集合中的标引词的数目为 n 时，模型可以推广到 n 维空间的欧几里得距离。根据线性向量模型理论，广义的析取提问和合取提问可以分别表示为

$$\boldsymbol{q}_{\mathrm{or}}=t_1 \vee^p t_2 \vee^p \cdots \vee^p t_n$$

$$\boldsymbol{q}_{\mathrm{and}}=t_1 \vee^p t_2 \vee^p \cdots \vee^p t_n$$

这里，p 是一个可变的量，$1 \leqslant p \leqslant \infty$ 的值在提问时就应当确定。则这两种文献-提问的相似度为

$$\mathrm{sim}(\boldsymbol{q}_{\mathrm{or}},\boldsymbol{d}_j)=\left[\frac{x_1^p+x_2^p+\cdots+x_n^p}{n}\right]^{\frac{1}{p}}$$

$$\mathrm{sim}(\boldsymbol{q}_{\mathrm{and}},\boldsymbol{d}_j)=1-\left[\frac{(1-x_1)^p+(1-x)^p+\cdots+(1-x_n)^p}{n}\right]^{\frac{1}{p}}$$

式中的 x_i 表示信息文档 $\boldsymbol{d}_j$ 中的第 i 个标引词的权值 $W_{i,j}$。由于 p 是一个变量，下面分析 p 的取值对相似度的影响。

(1) 当 $p=1$ 时，

$$\begin{aligned}\mathrm{sim}(\boldsymbol{q}_{\mathrm{and}},\boldsymbol{d}_j) &= 1-\frac{n-(x_1+x_2+\cdots+x_n)}{n}=\frac{x_1+x_2+\cdots+x_n}{n}\\ &=\mathrm{sim}(\boldsymbol{q}_{\mathrm{or}},\boldsymbol{d}_j)\end{aligned} \tag{5-11}$$

则布尔逻辑表达式中的布尔逻辑运算符"∧"、"∨"已毫无区别，两者的功能都减退为0，相似度的计算采取简单的向量空间模型余弦函数法，即

$$\mathrm{sim}(\boldsymbol{d}_j,\boldsymbol{q})=\frac{\overline{\boldsymbol{d}_j}\cdot\overline{\boldsymbol{q}}}{|\overline{\boldsymbol{d}_j}|\times|\overline{\boldsymbol{q}}|}=\frac{\sum_{i=1}^{t}W_{i,j}\times W_{i,q}}{\sqrt{\sum_{i=1}^{t}(W_{i,j})^2}\times\sqrt{\sum_{i=1}^{t}(W_{i,q})^2}} \tag{5-12}$$

(2) 当 $p=\infty$ 时，标引词的权值在[0,1]上，扩展布尔模型就变成建立在模糊逻辑上的布尔检索模型，则"信息文档-提问"之间的相似度为

$$\mathrm{sim}(\boldsymbol{q}_{\mathrm{or}},\boldsymbol{d}_j)=\lim_{p\to\infty}\left[\frac{x_1^p+x_2^p+\cdots+x_n^p}{n}\right]^{\frac{1}{p}}=\max(x_1,x_2,\cdots x_n)$$

$$\begin{aligned}\mathrm{sim}(\boldsymbol{q}_{\mathrm{and}},\boldsymbol{d}_j) &= \lim_{p\to\infty}\left\{1-\left[\frac{(1-x_1)^p+(1-x)^p+\cdots+(1-x_n)^p}{n}\right]^{\frac{1}{p}}\right\}\\ &= 1-\max(1-x_1),(1-x_2),\cdots,(1-x_n)\\ &= \min(x_1,x_2,\cdots,x_n)\end{aligned} \tag{5-13}$$

(3) 当 p 值在 1 与∞之间时，扩展布尔模型就介于向量模型和布尔模型之间，p 值越大，∧和∨的功能就越强；p 值越小，∧和∨的功能就越弱，直至 $p=1$，其功能完全消失。见图 5-5。

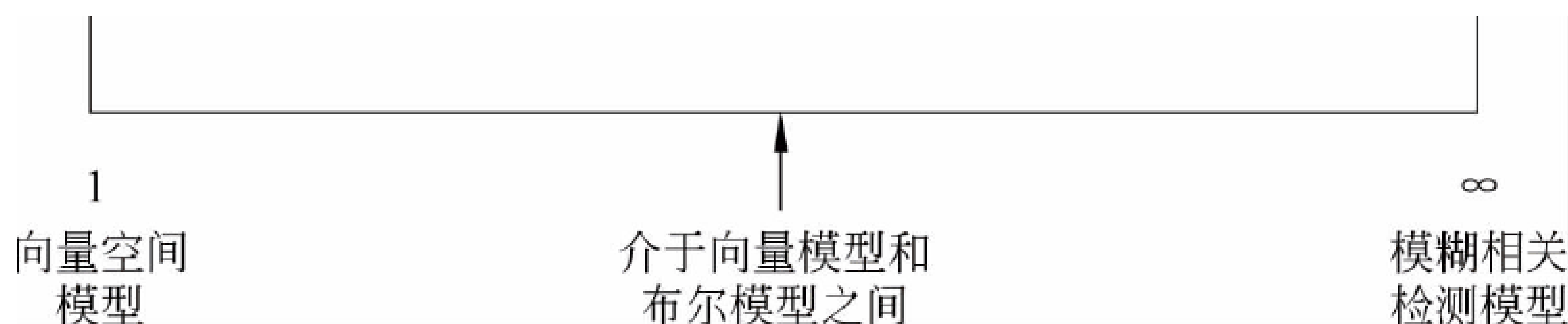


图 5-5　p 值的变化范围

对于提问语言的处理一般是按预先定义的次序对运算符进行分组而展开的，比如对于提问 $\boldsymbol{q}=(t_1\wedge^p t_2)\vee^p t_3$，信息文档 $\boldsymbol{d}_j$ 与提问 $\boldsymbol{q}$ 的相似度通常计算为

$$\mathrm{sim}(\boldsymbol{q},\boldsymbol{d}_j)=\left\{\frac{\left[1-\left(\frac{(1-x_1)^p+(1-x_2)^p}{2}\right)^{\frac{1}{p}}\right]^p+x_3^p}{2}\right\}^{\frac{1}{p}} \tag{5-14}$$

扩展布尔信息检索模型放宽了这种用代数学的距离来解释一元布尔运算，在某种意义上说，扩展的布尔检索模型是一个混合模型，它既有基于集合理论的信息检索布尔模型、信息检索模糊模型的特征，也具有基于代数理论的向量空间信息检索模型的特征，但人们通常倾向于把它归为集合论模型。

布尔模型和扩展的布尔模型主要是基于康托(Contor)的经典集合论：一个元素 a 和一个集合 A 的关系只存在 $a\in A$，$a\notin A$ 两种情况，经典集合论容不得模糊的概念，这对于信息检索过程中所存在的模糊性的解释造成一定的困难。检索中的模糊性主要体现在以下四个方面：

(1) 用户通常不能准确地说明他所需要的信息，在检索过程中会出现“全部”、“一些”等数量上的模糊关系和“相关”、“紧密相关”等相关性方面的模糊概念。

(2) 系统中所采用的信息文档标识只是信息文档内容的部分和不准确的表示。

(3) 大部分信息文档只是与用户提问部分相关。

(4) 用户对于检索结果的满意程度也具有不确定性。为了解决这种模糊性引起的不确定问题，人们引入模糊集合理论来构建模糊集合模型。

扩展布尔模型是常规布尔检索精确匹配的严格性和向量处理模式提问的无结构性的折中，它用代数距离的方式来解释并放松了布尔操作的限制要求，因而有效融合了传统的布尔、向量等检索模型的处理思想。扩展布尔模型的主要特点分析有以下几个方面：

(1) 与传统布尔检索中的倒排文档技术相兼容，支持使用标准布尔逻辑表达的提问式结构。

(2) 允许在文档和提问式中进行词加权处理；支持按相似度的大小排序输出检索结果。

(3) 通过调整参数 p 的取值，可以灵活选择并得到不同的检索结果。

(4) 扩展布尔逻辑检索模型适用于反馈信息系统。

(5) 可以对信息文档的标引词和提问词分别加权，以反映信息文档中词语的相对重要性程度和用户提问的侧重点。

5.4 信息检索代数模型

检索代数模型是以线性代数、矩阵计算等数学理论为基础，利用代数论基本知识揭示信息间关系的检索模型，它在信息检索的发展中发挥着重要作用。检索代数模型主要包括向量空间模型、隐含语义索引模型、神经网络模型等具体类型。

5.4.1 信息检索向量空间模型

1. 向量空间模型概述

Gerard Salton 在 20 世纪 60 年代提出了向量空间模型(vector space model，VSM)对信息特征进行表达，后来成功应用于很多文本检索系统(system for the manipulation and retrieval of text，SMART)，VSM 理论框架到现在仍然是信息检索研究的重要基础理论之一。但随着网络信息量的剧烈膨胀和网络信息格式的多样化，这种方法查询的结果往往会与用户真实的需求相差甚远，而且产生的无用信息量非常多，许多用户希望的个性化查询无法实现(个性化查询就是将一般的查询结果根据用户的个性模型进行二次检索，以适应用户个人的需求)，为此人们从许多方面对 VSM 进行优化和改进，以期获得更高的查询精度和效率。

2. 文档向量的构造

对于任一信息文档 $\boldsymbol{d}_j \in D$，我们可以把它表示为如下 t-维向量的形式：

$$\boldsymbol{d}_j = (w_{1j}, w_{2j}, \cdots, w_{tj}) \tag{5-15}$$

其中，向量分量 w_{ij} 代表第 i 个索引词 k_i 在文档 $\boldsymbol{d}_j$ 中所具有的权重，t 为系统中索引词的个数。在布尔模型中，w_{ij} 的取值范围是$\{0,1\}$；在向量空间模型中，由于采用“部分匹配”策略，w_{ij} 的取值范围则是一个连续的实数区间$[0,1]$。

众所周知，一篇文档信息中会标引出多个不同的索引词，而这些索引词对表达该篇文档信息主题的能力往往是不同的。换句话说，每个索引词应该具有不同的权值。如何计算文档向量中每个索引词的权值，不仅关系到文档向量的形成，也关系到后续的检索匹配结果。

目前，索引词权值计算方案有很多种。在进行加权计算时，索引词权值的大小主要依赖于对索引词的各种频率数据的统计，并通常考虑两个方面的因素：局部权值和全局权值。所谓“局部权值”，是指第 i 个索引词在第 j 篇文档中的权值，而“全局权值”则是指第 i 个索引词在整个系统文档集合中的权值。现在，假设 N 为检索系统文档总数，n_i 为系统

中含有索引词 k_i 的文档数，freq_{ij} 为索引词 k_i 在文档 $\boldsymbol{d}_j$ 中的出现次数，idf_i 表示索引词 k_i 的逆文档频率(inverse document frequency，idf 或 IDF)，maxtf_j 表示文档 $\boldsymbol{d}_j$ 中所有索引词出现次数的最大值，那么，对于文档 $\boldsymbol{d}_i$ 中索引词 k_i 的权值计算方法如下：

$$
\begin{aligned}
f_{ij} &= \mathrm{freq}_{ij}/\max \mathrm{tf}_j \quad &\text{(局部权值)}\\
\mathrm{idf}_i &= \log (N/n_i) \quad &\text{(全局权值)}\\
w_{ij} &= f_{ij} \times \mathrm{idf}_i \quad &\text{(索引词全值)}
\end{aligned}
\tag{5-16}
$$

式(5-16)是一种最为流行的权值计算公式，被研究人员称为“tf-idf(词频-逆文档频率)”加权模式。基于这一加权模式的计算公式还有一些，对于它们的加权效果，研究人员也进行了相当多的试验分析。

3. 提问向量的构造

在向量空间模型中，用户的信息需求被加工、转换为提问向量，并用与文档向量类似的表示形式表示，即

$$
\boldsymbol{q} = (w_{1q}, w_{2q}, \cdots, w_{tq}) \tag{5-17}
$$

这里，t 为系统索引词的总数，向量分量 w_{iq} 表示第 i 个索引词 k_i 在提问 q 中的权值，且有 $w_{iq} \geqslant 0$。至于如何评估 w_{iq} 的权值，一个推荐性的计算公式是

$$
w_{iq} = (0.5 + 0.5 \times \mathrm{freq}_{iq}/\mathrm{maxtf}_q) \times \log (N/n_i) \tag{5-18}
$$

其中，freq_{iq} 为在表述用户信息需求的文本内容中索引词 k_i 的出现次数，而 maxtf_q 则为在表述用户信息需求的文本信息中使用的所有索引词出现次数的最大值。

4. 匹配函数的选择及相似度阈值的确定

在文档与提问向量化表示的基础之上，文档与查询提问之间的相关程度(即相似度)就可以由它们各自向量在 t-维空间的相对位置来决定。一般地，相似度计算函数 $\mathrm{sim}(\boldsymbol{d}_j, q)$ 可以有非常多样化的选择，但较常采用的相似度计算指标是两个向量夹角的余弦函数(见图 5-6)。

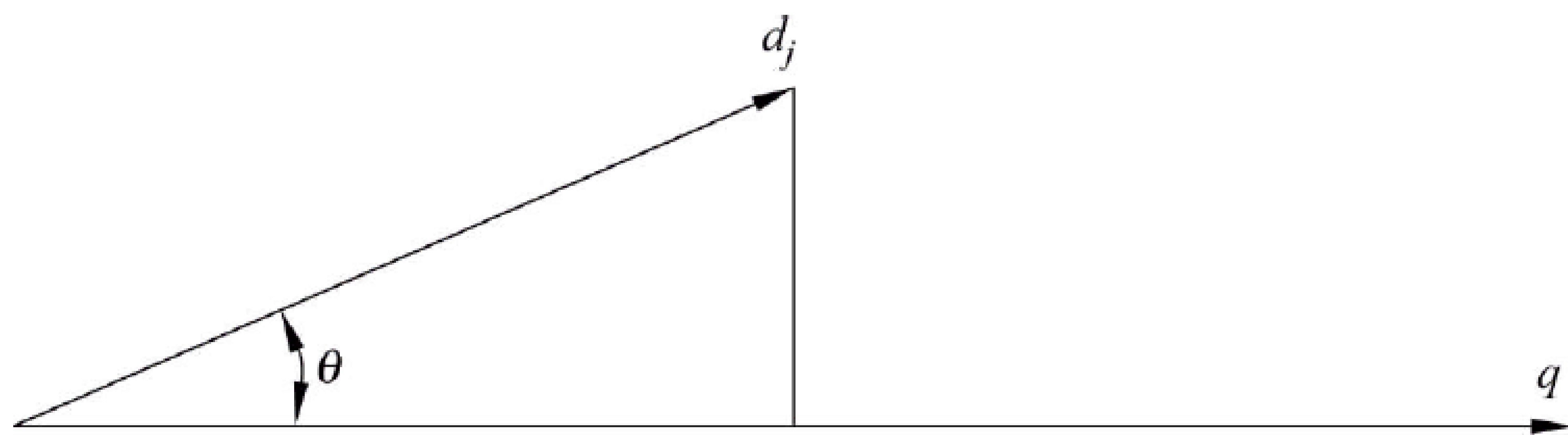


图 5-6　文档向量与提问向量的夹角及余弦值

按照两个向量夹角余弦的计算含义，文档 $\boldsymbol{d}_j$ 和提问 q 的相似度值就可以通过下面的

计算公式获得：

$$\mathrm{sim}(\boldsymbol{d}_j,\boldsymbol{q}) = (\boldsymbol{d}_j \cdot \boldsymbol{q})/(|\ \boldsymbol{d}_j\ |\times|\ (q)\ |) = \frac{\sum_{i=1}^{t} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \times \sqrt{\sum_{i=1}^{t} w_{iq}^2}} \qquad (5\text{-}19)$$

式(5-19)中，$|\ \boldsymbol{d}_j\ |$ 和 $|\ \boldsymbol{q}\ |$ 分别表示文档向量 $\boldsymbol{d}_j$ 和提问向量 $\boldsymbol{q}$ 的模(norm)或长度，分子 $\boldsymbol{d}_j \cdot \boldsymbol{q}$ 是两向量的内积。由于 $w_{ij} \geqslant 0$ 和 $w_{iq} \geqslant 0$，因此有 $0 \leqslant \mathrm{sim}(\boldsymbol{d}_j,\boldsymbol{q}) \leqslant 1$。这样一来，检索处理不仅能判断文档是相关还是不相关，而且还可以定量化地判断系统所有文档与某一提问的相关度大小，并能够按照其相关度值的降序排列方式输出命中的结果文档。

为更有效地得到一个合理的检索结果，需要进一步指定一个相关度阈值(threshold)λ，凡与提问向量的相关度值大于 λ 的文档，都将作为检索结果提供给用户。如此，向量空间模型的检索匹配便有一种"部分匹配"策略思想。

5. 基于向量空间的信息检索描述

一个向量空间是由一组线性无关的基本向量组成，向量维数与向量空间维数一致，并可以通过向量空间进行描述。向量空间模型描述如下：

概念 1：文档 D(document)，泛指文档或文档中的一个片段(如文档中的标题、摘要、正文等)。

概念 2：特征项 t(term)，指出现在文档中能够代表文档性质的基本语言单位(如词语等)，也就是通常所指的检索词。这样一个文档 D 就可以表示为 $\boldsymbol{D}(t_1,t_2,\cdots,t_n)$，其中 n 就代表了检索字的数量。

概念 3：特征项权重 W_k(term weight)指特征项 t_n 能够代表文档 $\boldsymbol{D}$ 能力的大小，体现了特征项在文档中的重要程度。这样文档 $\boldsymbol{D}$ 的向量可以表示为 $\boldsymbol{D}(\omega_{n1},\omega_{n2},\cdots,\omega_{nm})$，其中 $\omega_1,\omega_2,\omega_m$ 分别代表文档 $\boldsymbol{D}$ 特征项 $t_1,t_2,\cdots,t_n$ 的特征项权重。在网络索引文件中，每一个向量对应一个 URL，当用户检索查询一个文档内容时，如果匹配，则向量 $\boldsymbol{D}$ 对应的特征项 t 值为 1，否则值为 0，如下所示：

Term ID	T_1	T_2	…	T_n
D_1	0	1	…	0
D_2	0	1	…	0
…	…			

$$\text{查询向量}\ \boldsymbol{q}_j\begin{cases}1, & \text{若}\ t_i \in \text{查询条件 QS}\\ 0, & \text{若}\ t_i \notin \text{查询条件 QS}\end{cases}$$

概念 4：相似度 S(similarity)，指两个文档内容相关程度的大小，当文档以向量来表示时，可以使用向量文档向量间的距离来衡量，一般使用内积或夹角 θ 的余弦来计算，两者夹角越小说明相似度越高。由于查询也可以在同一空间里表示为一个查询向量（见图 5-7），可以通过相似度计算公式计算出每个文档向量与查询向量的相似度，排序这个结果后与设立的阈值进行比较。如果大于阈值，则网页与查询相关，保留该页面查询结果；如果小于，则不相关，过滤此网页。这样就可以控制查询结果的数量，加快查询速度。

$$\mathrm{sim}(\boldsymbol{D}_1,\boldsymbol{D}_2) = \sum_{K=1}^{n} W_{1k} \times W_{2k} \tag{5-20}$$

$$\mathrm{sim}(\boldsymbol{D}_1,\boldsymbol{D}_2) = \cos\theta = \frac{\sum_{k=1}^{n} W_{1k} \times W_{2k}}{\sqrt{\left[\left(\sum_{k=1}^{n} W_{1k}^2\right)\left(\sum_{k=1}^{n} W_{2k}^2\right)\right]}} \tag{5-21}$$

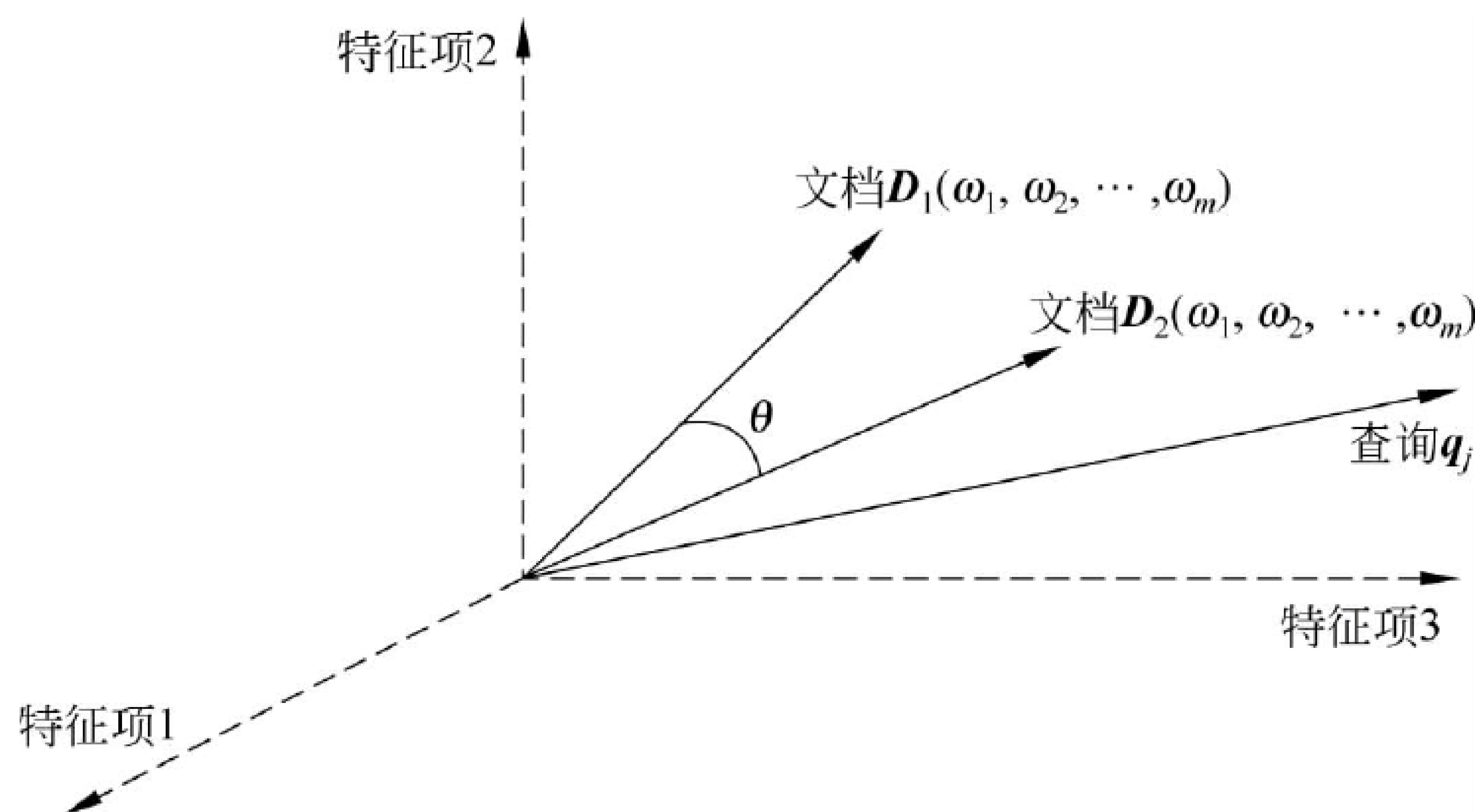


图 5-7　文档 VSM 及相似度 Sim($\boldsymbol{D}_1$、$\boldsymbol{D}_2$)

6. 信息检索向量空间数学模型工作机制

向量空间模型是目前信息检索最常用的数学模型之一，在 WWW 信息方面，向量空间模型比布尔模型等传统模型更合适。基于向量空间模型的信息检索一般过程是：①将各个文档和查询都表示成为向量；②计算查询与各个文档之间的相似度；③按照查询与各个文档之间的相关度对相关的文档进行排序；④将排序后的文档以线性列表的形式返

回给用户。

根据上述知识可以引出如图 5-8 所示的向量空间信息检索模型机制图,这里需要解决特征项的生成和加权、相似度的计算(检索运算)等一系列问题。由于向量检索中采用向量间的某种距离度量来反映文档对的满足程度,所以相似度的值最好能与真实情况相符。而且计算简便,计算出的值最好能归一化到[0,1]区间上,分布尽可以均匀,使阈值的选择容易一些。直接选定相似度阈值的办法有时不太好控制,这时可以根据相似度对文档排序并直接给定输出的文档数目。

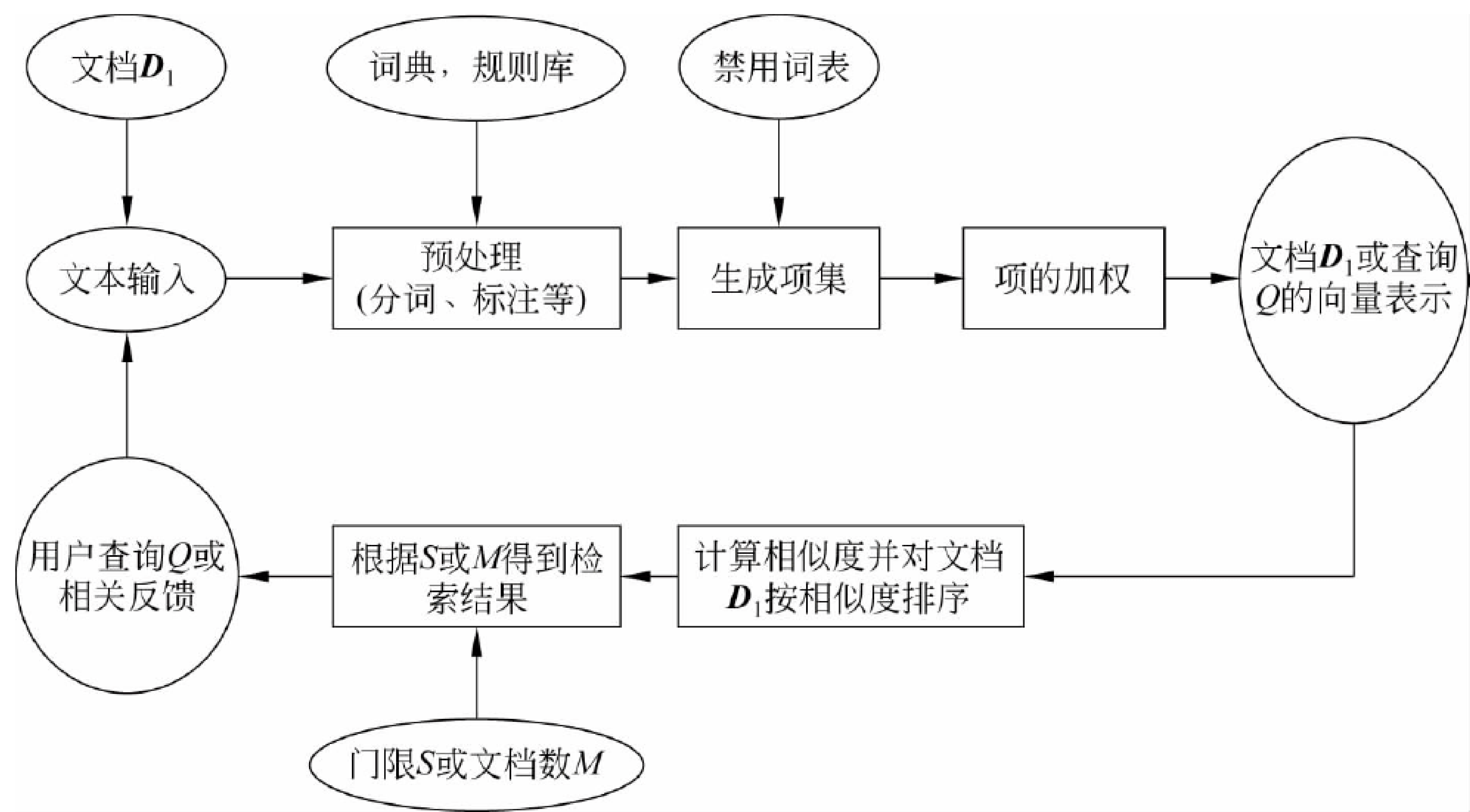


图 5-8 向量空间信息检索模型机制

7. 向量空间信息检索模型的不足

从向量空间模型的特点可以看出,在特征项确定的情况下,特征项的权重计算是文档分类的关键,特征项权重计算常用的方法有布尔函数、开根号函数、对数函数、TFIDF 函数等。其中 TFIDF 函数应用最为广泛,基本思路是使用频率因子 TF(term frequency)进行特征项的赋权,同时还要考虑文档集因子 IDF(inverse document frequency),体现出查询内容与文档的相关度大小,一般采用使用出现频率的倒数来计算,但是 TFIDF 函数也存在缺点,它虽然考虑了出现特征项的文本在整个文档集中的比例,却不能很好地把握特征项在文本集合中分布的差异,所以影响了分类的最终效果。

VSM 的第一个问题是由于特征项在文档中的不同位置代表不同的权重,而不同的关键词长度也会影响权重的大小。例如"汽车修理"一词在查询时,如果该词出现在文档的

标题处,则其权重一定比出现在文章的摘要中要高,而出现在摘要中的权重一定要比出现在正文中要高;而且如果文档 $\boldsymbol{D}_1$ 的长度比文档 $\boldsymbol{D}_2$ 长,那么在 $\boldsymbol{D}_2$ 中的权重也应该比 $\boldsymbol{D}_1$ 要高,其相似度也应该大一些。对于中文文档,关键词的长度越长,则在文档中出现的概率就越小,所以较长的关键词要比较短的包含更多的信息。在实际情况中,如果同一特征项在不同文档中出现的次数不同,那么在出现频率较高的文档中,其权重应该较高(而不应该是统一权重值“1”)。在传统的 TFIDF 函数中,每增加一个文档都要重新计算向量,导致查询速度降低,同时由于使用频率因子,在扩大查询范围时,不可避免地会影响到查询的准确性。

VSM 的另一个问题在于查询和文档向量间是依靠链接来判断的,而且判断的依据是简单的两者相同关键词的比较。但实际情况是大量的关键词具有相同的语义,同一关键词也会有多种语义的解释描述(即产生了语义分歧)。例如“检索”一词,也可以是“查找”、“查询”等,对用户来说所指的含义可能是一个意思,但在 VSM 中这几个词是完全不同的概念,也就是说用户使用“检索”这个关键词去查询时,包含相关的“查找”、“查询”的文档会检索不出,而另一方面,可能许多不相关的文档反而会被检索出来。

8. 改进的 VSM 方法

传统的 VSM 主要的缺陷就是特征项相互独立与自然语言多样性有矛盾。实际上主要考虑两个方面的改进:一个是检索关键词的长度和出现在文档中的位置对权重的影响,另一个就是要考虑检索关键词的语义环境影响。

1) 加权的 VSM 改进算法

$$W_i = \lambda \times \mathrm{tf}_i \times \log\left(\frac{N}{n_i} + 0.1\right) + \frac{\mathrm{tf}_i}{l_i} \tag{5-22}$$

其中 λ 为位置加权系数,表示检索文本在文档不同位置的加权处理参数,按照检索文本在文档中的位置不同,一般分为标题、摘要、关键词、正文、结论和超链接六个位置,分别赋予不同的加权系数,由于 Web 文档信息都是通过链接来完成的,Web 上的各种标记和链接包含了页面的结构信息,应该给予足够的重视和利用。例如,在链接 $r \rightarrow s$ 中,r 的连接标记若为文档 $\boldsymbol{D}_1$〈a href=“http: www. china…”〉锚文本〈/a〉文档 $\boldsymbol{D}_2$,其中锚文本对目标 URL=“http: www. china…”会有比较准确的描述,而文档 $\boldsymbol{D}_1$、$\boldsymbol{D}_2$ 就次之,所以对于出现在锚文本和文档 $\boldsymbol{D}_1$、$\boldsymbol{D}_2$ 中的每一个特征项应赋予较高的权重系数。

另外一个关键的加权位置在一些语义的重点语句位置,如“综上所述”、“结束语”、“主要在于”等关键语句中,其值可以从辅助检索词表中获取。一般位置加权系数 λ 的计算可

以考虑使用各分部分的频率与不同位置加权系数的乘积和来表示。

$$\lambda = tf_0 + tf_1 \times \lambda_1 + tf_2 \times \lambda_2 + tf_3 \times \lambda_3 + tf_4 \times \lambda_4 + tf_5 \times \lambda_5 \quad (5\text{-}23)$$

其中 tf_0 为对正文关键词统计的词频数；tf_1，tf_2，tf_3，tf_4，tf_5 分别为标题、摘要、关键词、结论、超链接中的词频；λ_1，λ_2，λ_3，λ_4，λ_5 分别为其加权系数。

tf_i 为特征项频率；N 为总文档数量；n_j 为包含特征项 W_i 的文档数；l_i 为文档长度，使用$\frac{tf_i}{l_i}$来表示文本能够代表文档内容的能力。例如，虽然“计算机”一词出现在文档标题和正文中的频率相同，但由于标题比正文文档长度要小得多，所以我们认为“计算机”一词在标题中的权重要比在正文中的权重大得多。

2）辅助检索词表和个性化协同检索设计

由于自然语言的特点，从语法角度来看，许多关键词的含义只起修饰的作用（如形容词、副词），并不能表示独立的概念，这些带有修饰和限制性的词在很大程度上代表了用户查询的需求，如果忽略这部分内容，将会产生许多不相关的查询结果；同时由于一词多义、一义多词的现实情况，简单地以检测一个文档与查询语句间的特征项是否相同来判断是否具有相关性，会使许多真正与之相关的文档反而没有被检索出来。

因此需要设计一个辅助检索词表，用来存储同义词和修饰限制词语，借助这个数据库，将用户查询的特征值进行语义扩展，将检索关键词与字典库中的同义词和修饰词结合起来，形成新的检索特征项，这样就将孤立的用户初始检索词变成了一个具有自然语义的检索词，在查询时就可以将只含有初始检索词而不能表示辅助检索词表修饰语的文档过滤，从而提高检索精度和效率。

另一方面，利用“个性化信息库”来分析用户兴趣，并根据以往用户的检索信息内容推荐早期用户兴趣，配合概念检索进行协同，以期获得更为个性化的信息服务；同时将每一次的检索结果、用户兴趣等进行信息反馈、定期刷新，不断充实改进“个性化信息库”；此外，不同的用户对相同的检索内容会有着不同的理解和期望结果，所以还可以根据“个性化信息库”来设计不同的检索结果库，以期得到个性化的检索结果。

在个性化协同算法中，可以将用户（user）模型以向量形式来表示：$\boldsymbol{U}=(u_1, u_2, \cdots, u_n)$，其相似度可以用与特征向量的内积来计算：$\mathrm{Sim}(\boldsymbol{U}, \boldsymbol{W}) = \sum_{k=1} u_k \times \omega_k$，即计算用户模型与文档特征项的相似度，排除非用户感兴趣的页面，从而实现查准率的提高。改进后的 VSM 流程如图 5-9 所示。

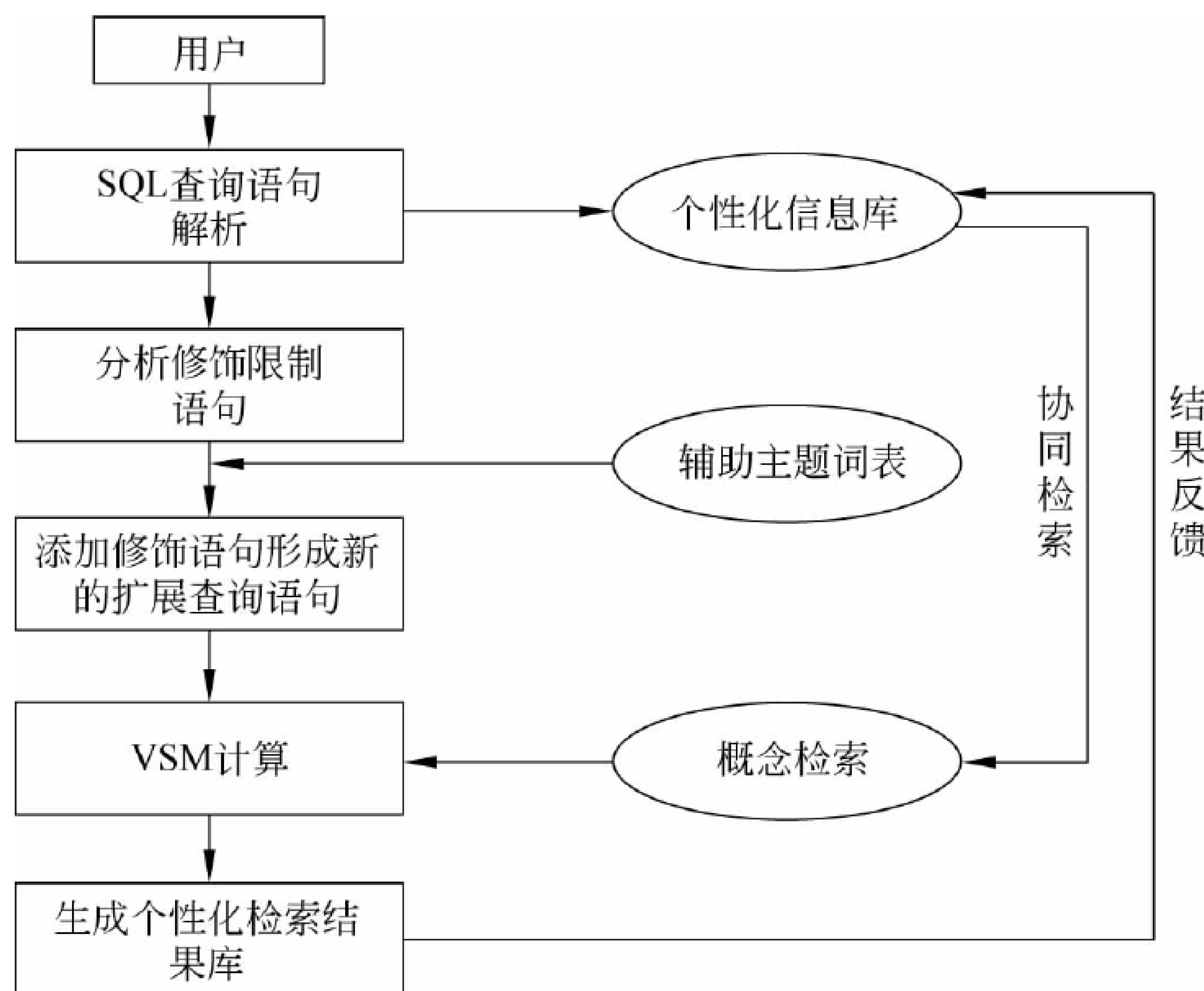


图 5-9　VSM 改进算法流程图

5.4.2　潜在语义索引模型

1. 潜在语义索引的提出

传统的分类模型一般是用词条作为特征的，为了降低检索系统的复杂度，一般认为检索词与检索词之间是相互独立的，这显然是与事实相违背的，因此向量空间模型的效果一直有不足之处。自然语言中词语的多义性(polysemy)与同义性(synonymy)现象普遍存在，当初萨尔顿等人在 VSM 中关于特征项(即索引词)之间相互独立的基本假设(正交假设)，在实际检索的信息处理过程中很难满足信息获取需要。那么，如何修正“正交假设”的缺陷与不合理性，并将文本检索处理水平从离散的索引词形式匹配深入到概念或语义匹配的层次上，成为代数检索迫切需要考虑的问题。从 20 世纪 80 年代末开始，杜麦斯(S. T. Dumais)、贝瑞(M. W. Berry)等研究人员基于 VSM 理论框架，分析并提出了一种新的信息检索模型——潜在语义索引或隐含语义索引(latent semantic indexing，LSI)。在有些研究文献中，研究人员也称为潜在(或隐含)语义分析(latent semantic analysis，LSA)。

在用词条来表示文本的时候，大量存在的同义词、近义词和多义词，使得特征之间相

互独立的假设不能成立。LSI 通过统计大量文本中这些词的共现信息来发掘它们的内部联系，称为文本的语义。LSI 认为每个文档都包含有几种语义，这些语义之间是相互独立的，如果可以用这些语义来表示文档，并拿它们来进行计算，则在降低计算复杂度的同时，还可以保持很好的效果。由于这种语义不能直接得到，只能通过对文档特征的分析得到，它是潜藏在文档信息特征之间的，所以称为“潜在语义”。

潜在语义索引可以看成是一种扩展的向量空间模型，用于发现文本信息中的语义关系。潜在语义索引基于如下假设：文档中的词条与词条之间是存在一定关联的，只不过潜在的这个语义被文档中词条的语义和形式上的多样性掩盖得不明显而已。LSI 能够加强相关词条（或文档）之间的关联性，而削弱非相关词汇（或文档）之间的关联性，将高维空间中的文档向量（或词条向量）投影到低维的潜在语义空间中，使得原来没有任何共同项的两个文档（或词汇）经过 LSI 处理后有可能找到彼此间比较有意义的关联性，体现文档（或词汇）间的语义。

2. 潜在语义索引的基本思想

潜在语义索引使用了向量空间模型的方法来表示“词汇-文本”矩阵，是对向量空间模型的扩展，其中每一行代表一个词汇向量，每一列代表文本集中的一个基于关键词的向量空间模型（VSM），用 $A=\{a_{ij}\}_{m\times n}$ 表示 m 个词汇和 n 个文本构成的文本集合，它的优点在于将非结构化的文本表示为向量形式，使得各种信息检索的基本数学处理成为可能。但是，向量空间模型是基于词汇之间关系相互独立的基本假设（正交假设），在实际情况下很难得到信息查询的需求满足，文本中出现的词往往存在一定的相关性，在某种程度上会影响计算结果。

LSI 则将自然语言中的每个文本视为以词汇为维度的空间中的一个点，认为一个包含语义的文本出现在这种空间中，它的分布绝对不是随机的，而是服从某种语义结构的。同样地，也将每个词汇视为以文本为维度的空间中的一个点。文本是由词汇组成的，而词汇又要放到文本中去理解，体现了一种“词汇-文本”双重概率关系。

LSI 把词汇中的一些不经常的用法，如一些词汇的误用，或不相关的词汇偶然出现在一起，还有高频词、低频词等不能代表文本主题的词汇视为“噪声”，应当从主要语义结构中排除掉。利用截断的奇异值分解降维的方法，达到信息过滤和去除噪声的目的。通过对“词汇-文本”矩阵 $\boldsymbol{A}$ 进行截断的奇异值分解，得到矩阵 $\boldsymbol{A}$ 的序为 k 的“近似矩阵”，从数据压缩的角度看，“近似矩阵”是序为 k 的前提下矩阵 $\boldsymbol{A}$ 的最小二阶意义上的最佳近似。LSI 不同于 VSM 中文本和词汇的高维表示，而是将文本和词汇的高维表示投影在低维的潜在语义空间中，缩小了问题的规模，得到词汇和文本的不再稀疏的低维表示，同时这种

低维表示揭示出了“词汇-文本”之间语义上的联系。

3. 潜在语义索引的数学基础

实验表明：潜在语义索引通过奇异值分解，不仅减少了“词汇-文本”矩阵的维数，而且大大消减了一直困扰基于关键词的信息检索的文本中词汇的同义性和多义性问题，那么，潜在语义索引的数学依据是什么呢？通过两个关于奇异值分解定理来进行剖析。

定理 1：假设 $\boldsymbol{A}$ 的奇异值分解由公式给出，并且有

$$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_r \geqslant \lambda_{r+1} = \cdots = 0$$

$R(\boldsymbol{A})$和 $N(\boldsymbol{A})$分别表示 $\boldsymbol{A}$ 的表示区域和 $\boldsymbol{A}$ 的零空间，则有

（1）阶特性：$\mathrm{rank}(\boldsymbol{A})=r, N(\boldsymbol{A})=\{\boldsymbol{v}_{r+1},\cdots,\boldsymbol{v}_n\}, R(\boldsymbol{A})=\mathrm{span}\{\boldsymbol{u}_1,\cdots,\boldsymbol{u}_r\}$，

$$U=[\boldsymbol{u}_1,\cdots,\boldsymbol{u}_m],\quad V=[\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_n]$$

（2）二阶分解性：

$$\boldsymbol{A}=\sum_{i=1}\boldsymbol{u}_i\cdot\lambda_i\cdot\boldsymbol{v}_i^{\mathrm{T}} \tag{5-24}$$

（3）规范性：$\|\boldsymbol{A}\|_F^2=\lambda_1^2+\lambda_2^2+\cdots+\lambda_r^2$，　$\|\boldsymbol{A}\|_2=\lambda_1$

其中，$\|\cdot\|_F^2$ 和 $\|\cdot\|$ 分别代表矩阵的 F-范数和谱范数，定理 1 说明了单位向量 $\boldsymbol{u}_i,\boldsymbol{v}_i$ 与矩阵 $\boldsymbol{A}$ 的关系，同时也体现了矩阵 $\boldsymbol{A}$ 的特征值与其范数的关系。

但是，向量 $\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_r$ 对“词汇-文本”矩阵 $\boldsymbol{A}$ 的影响程度是不一样的。因此，常常需要对矩阵 $\boldsymbol{A}$ 相应的语义空间进行压缩，由于 r 个特征值是按大小排序的，只保留前 k 个最大的特征值，即所谓的对 $\boldsymbol{A}$ 进行奇异值分解。

所以上面最重要的是奇异值分解的阶的特性，它表明可以将矩阵的奇异值作为矩阵定性分析的定量手段。而奇异值分解的二阶分解性表明，在很多应用场合中可以对矩阵进行大胆的压缩。

定理 1 的三个方面可以用来证明下列定理。

定理 2：假设 $\boldsymbol{A}$ 的奇异值分解由公式给出，$r=\mathrm{Rank}(\boldsymbol{A})\leqslant p=\min(m,n)$，对于任意的 $k\leqslant r$，定义：

$$\boldsymbol{A}_k=\sum_{i=1}^{k}\boldsymbol{u}_i\cdot\lambda_i\cdot\boldsymbol{v}_i^{\mathrm{T}} \tag{5-25}$$

那么，

$$\min_{r(\boldsymbol{B})=k}\|\boldsymbol{A}-\boldsymbol{B}\|_F^2=\|\boldsymbol{A}-\boldsymbol{A}_k\|_F^2=\lambda_{k+1}^2+\cdots+\lambda_p^2; \tag{5-26}$$

$$\min_{r(\boldsymbol{B})=k}\|\boldsymbol{A}-\boldsymbol{B}\|_2=\|\boldsymbol{A}-\boldsymbol{A}_k\|_2=\lambda_{k+1} \tag{5-27}$$

这一重要结论表明，由 $\boldsymbol{A}$ 的 k 个最大的奇异三元组构成的 $\boldsymbol{A}_k$ 是和 $\boldsymbol{A}$ 最接近的 k 序矩

阵，换言之，LSI 将“词汇-文本”矩阵从高序投影到低序后，尽可能地保留了原始矩阵 **A** 的大部分信息含量和查询能力。但是，这还不足以说明为什么 LSI 模型改进了查询能力。为此在一个比较严格的前提下，得到了下面的一个定理，这个定理能够更加明确地指出模型确实能够改进信息检索性能。

定理 3：假设 C 为一个纯粹的模型，ε 可分为包含 k 个主题的文本库模型，而且每一个词汇在某一主题中出现的概率最大为 τ，τ 为一个大于 0 的足够小的值。若有 m 个文本由 C 模型产生，则序为 k 的 LSI 以$(1-0()m^{-1})$的概率 $0(\varepsilon)$偏向 C。

4. 潜在语义索引的特点

与传统的向量空间模型相比，LSI 的优点在于以下几方面：

(1) 利用潜在的语义结构表示词汇和文本，将词汇和文本映射到同一个 k 维的语义空间内，向量的含义发生了很大变化。它反映的不再是简单的词汇出现频率和分布关系，而是强化的语义关系。在保持了原始大部分信息的同时，克服了传统向量空间表示方法产生的多义词、同义词和单词依赖的现象。同时，在新的语义空间中进行相似度分析，比使用原始的特征向量具有更好的效果，因为它是基于语义层而不仅仅是词汇层。

(2) 词汇和文本在相同的空间使得 LSI 更具灵活性，允许用户使用自然语言提交查询请求，查询条件可以是独立的词汇，也可以是文本信息内容，使得查询和反馈更容易。

(3) 用低维的“词汇-文本”关联空间代替了原来的“词汇-文本”独立空间，可以有效地处理大规模的文本集，有效地提高了检索的效率和准确性。

(4) LSI 不同于传统的自然语言处理过程和人工智能程序，它是完全自动的。所谓自动，就是 LSI 不需要人工干预，不需要预先具有语言学或者具备相似性知识，不使用人为构造的字典、知识基础、语义网络、文法、词法、句法剖析器等，它的输入只是原始的未经处理的文本序列。它完全是根据普通数学学习方法或机器学习方法，提取合适的维度语义空间，结合其他信息检索理论，达到有效展示对象和文本内容的语义关系目的。通过对大量的文本分析，LSI 可以自动地模拟人类的知识获取能力，甚至分类、预测的能力。

潜在语义索引模式以其数学理论严谨、处理文本信息过程思路清晰得到了信息检索领域的重视，该方法在语言建模、视频检索等方面取得了较为成功的应用，在朴素贝叶斯分类模型、KNN 模型和 VSM 模型中都被证明是非常有效的方法。但是，该方法也存在着一些不足之处：

(1) 潜在语义在进行信息提取时，忽视了词汇的语法信息甚至词汇出现的顺序性，它仍然是一种 Bag of word(词汇包)方法，即简单地通过所有词汇向量的线性拟合来产生文

本向量，表示文本的含义。但是句子的语法结构包含了词汇之间更深层次的语义关联信息，忽视这种关联信息在一定程度上影响了潜在语义对文本内容的准确性把握，虽然潜在语义通过新的空间在一定程度上实现了降维。

(2) 因子 k 值的选取直接关系到语义空间模型的效率，k 值过小则会使一些有用的信息丢失，k 值过大则会使运算复杂量增加，但是 k 值是一个可变的参数，对其确定是很困难的，现在还没有特别好的办法来解决。在实际中，人们一般只能通过反复的实验来确定这个值。

(3) 奇异值分解对存储空间的要求很大，运算的时间复杂度很高。SVD(语义向量划分)算法的时间代价是 $O(N^2k^3)$，N 是单词数和文本数的乘积，N 随文本数和单词数的增加而迅速增加，所以 SVD 不太适合动态变化的文本集。

5.4.3　神经网络检索模型

20 世纪 80 年代以来，人工神经网络研究取得重大进展，有关理论和方法已经发展成为一个介于数学、计算机科学、物理学、神经生理学等学科之间活跃的交叉研究领域。作为一种高度并行的信息处理方法，神经网络模型模拟人类脑神经系统的结构与功能，并以一种独特的方式对许多具有重大理论及实际意义问题的解决取得了突破性进展。

1. 神经网络研究概述

神经网络是指由大量神经元相互连接在一起所组成的神经结构，把神经元之间相互作用的关系进行数学模型化就可以得到神经网络模型。因此，神经网络模型主要来源于对人脑神经系统结构与功能的模拟，无论是单个神经细胞(或神经元)，还是神经网络的构成与作用方式。

研究表明，人脑是由约 10^{11} 量级个神经元构成的，而每一个神经元可以看做是一个基本的初等信号处理器。在一个神经元中，有信号的输入通道(即树突)和信号的输出通道(即轴突)。当信号从一个神经元经过连接通道(即突触)传递到另一个神经元时，一个相当复杂的生物物理及生物化学过程发生了，并可能产生两种不同的效果：接受信号的神经元或者被激发，或者被抑制。处于激发态的神经元又会产生新的脉冲信号，传向处于下游的每一个与之相连的神经元，并引起下游神经元不同的激发与抑制反应；而处于抑制态的神经元则不产生任何脉冲输出。上述信号传递与处理过程在整个神经系统的相关神经元之间不断重复进行，形成了人脑细胞的信号传播激活机制，并最终表现为：接受并处理输入信号，然后做出各种肢体或情绪上的反应。

值得注意的是，不同神经元之间的连接强度是不同的，而且连接强度也不是一成不变

的。通常连接或作用强度会随其激发与抑制行为的相关性时间的平均值成正比,这表明神经系统具有某种可塑性。

人工神经网络(artificial neural networks,ANN),被称为神经网络,是人工智能研究的一个重要领域。它是一个数学模型,通过模仿动物行为特征进行神经网络模型算法分析,并进行信息处理。通过调整节点之间的连接数目来实现处理信息的目的。

神经网络的工作过程分为两个方面,首先是训练期(也叫做学习期),通过测试信号的指导(有监督的情况)来训练样本,根据训练样本不断调整网络中边的连接权值。训练期之后是工作期,在此期间,神经网络的各个连接边的权值保持不变,而对测试样本进行输入计算,以实现对测试样本的打分。

一种常用的神经网络模型是基于反向传播学习算法(back propagation learning algorithm,BP),它的训练包括两个过程,包括正向传播和反向传播。在正向传播过程中,信息从输入层向输出层传播,中间可能经过零层到多层。在输出层根据实际输出和期望输出进行比较,得到误差。反向传播过程则是把输出层的误差信息从输出层逐层地往回传播,利用误差信息调整各连接边的权值,使得误差信息变小,以此达到训练目的。

2. BP 神经元

图 5-10 给出了第 j 个基本 BP 神经元,它模仿了生物神经元所具有的最核心也是最基础的功能:加权、求和和转移。其中 $x_1, x_2, \cdots, x_i, \cdots, x_n$ 分别代表来自神经元 $1, 2, \cdots, i, \cdots, n$ 的输入;$w_{j1}, w_{j2}, \cdots, w_{ji}, \cdots, w_{jn}$ 则分别表示神经元 $1, 2, \cdots, i, \cdots, n$ 与第 j 个神经元的连接强度,即权值;b_j 为阈值;$f(\cdot)$为传递函数;y_j 为第 j 个神经元的输出。

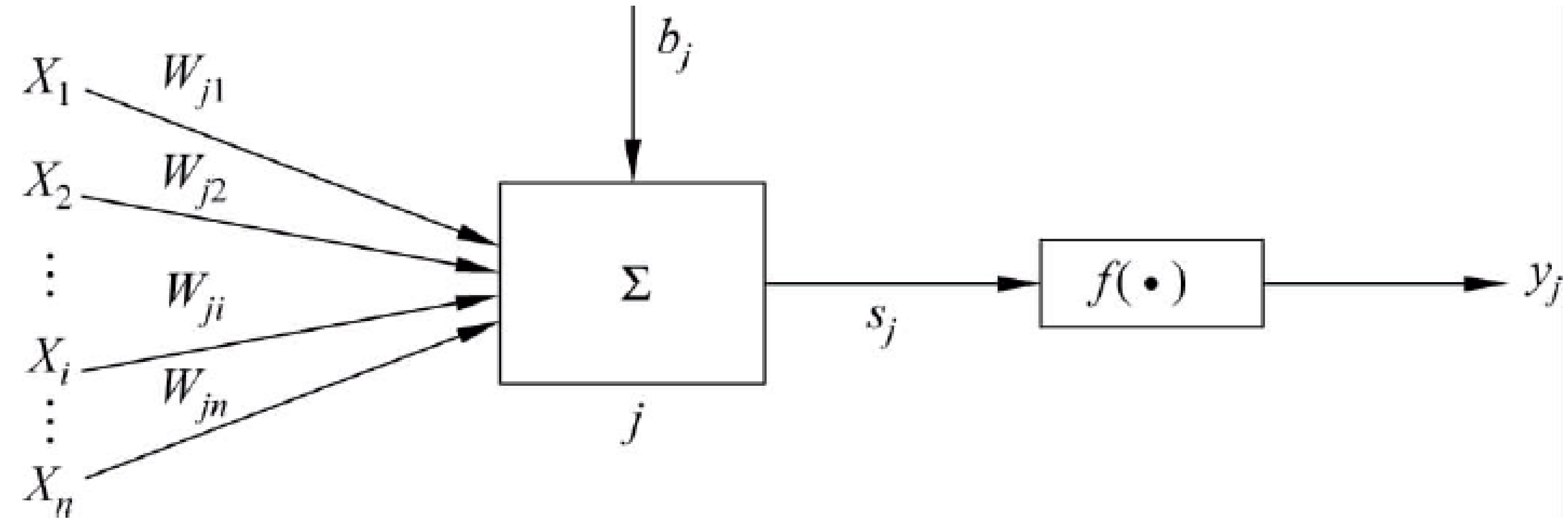


图 5-10　BP 神经元

第 j 个神经元的净输入值 S_j 为

$$S_j = \sum_{i=1}^{n} w_{ji} \times x_i + b_j = W_j \times X + b_j \tag{5-28}$$

其中:$X = [x_1 x_2 \cdots x_i \cdots x_n]^T$,$W_j = [w_{j1} w_{j2} \cdots w_{ji} \cdots w_{jn}]$,若视 $x_0 = 1$,$w_{j0} = b_j$,即令 X 及

W_j 包括 x_0 及 w_{j0}，则

$$X = [x_0 x_1 x_2 \cdots x_i \cdots x_n]^{\mathrm{T}}, \quad W_j = [w_{j0} w_{j1} w_{j2} \cdots w_{ji} \cdots w_{jn}]$$

于是节点 j 的净输入值 S_j 可表示为

$$S_i = \sum_{i=0}^{n} w_{ji} \times x_i = W_i \times X \tag{5-29}$$

净输入 S_i 通过传递函数(transfer function) $f(\cdot)$后，便得到第 j 个神经元的输出 y_1：

$$y_1 = f(s_j) = f\left(\sum_{i=0}^{n} w_{ji} \times x_i\right) = F(W_i \times X) \tag{5-30}$$

式中 $f(\cdot)$是单调上升函数，而且必须是有界函数，因为细胞传递的信号不可能无限增加，必有一最大值。

3. 神经网络基本原理

BP 神经网络算法是由数据流的向前计算(正向传播)和误差信号的反向传播两个过程构成。正向传播时，传播方向由输入层，经隐层传输到输出层，每层神经元的状态变化会影响到下一层神经元的状态值。如果在输出层得到的不是预期输出值，则将误差值反馈给前一层神经元，进行反向传播流程。通过这两个过程反复交替进行，使权向量空间误差函数梯度逐渐下降，动态迭代寻找最优权向量，使网络误差函数达到最小值，从而完成神经网络的学习过程。

1) 正向传播

设 BP 神经网络的输入层存在 n 个节点，隐层存在 q 个节点，输出层存在 m 个节点，输入层与隐层之间的权值分别为 v_{k1}，隐层与输出层之间的权值分别为 w_{jk}，如图 5-11 所示。隐层的传递函数为 $f_1(\cdot)$，输出层的传递函数为 $f_2(\cdot)$，根据 BP 神经网络算法原理，得隐层节点的 p 输出为

$$z_x = f_1\left(\sum_{i=0}^{n} v_{ki} \times x_i\right) \quad k = 1,2,\cdots,q \tag{5-31}$$

输出层节点的输出值为

$$y_i = f_2\left(\sum_{k=0}^{q} w_{jk} \times z_k\right) \quad j = 1,2,\cdots,m \tag{5-32}$$

通过上述原理，BP 神经网络则完成了 n 维空间向量对 m 维空间向量的映射。

2) 反向传播

(1) 定义误差函数

输入学习样本数 P 个，分别用 $x^1, x^2, \cdots, x^p$ 进行表示，假设第 p 个学习样本输入到

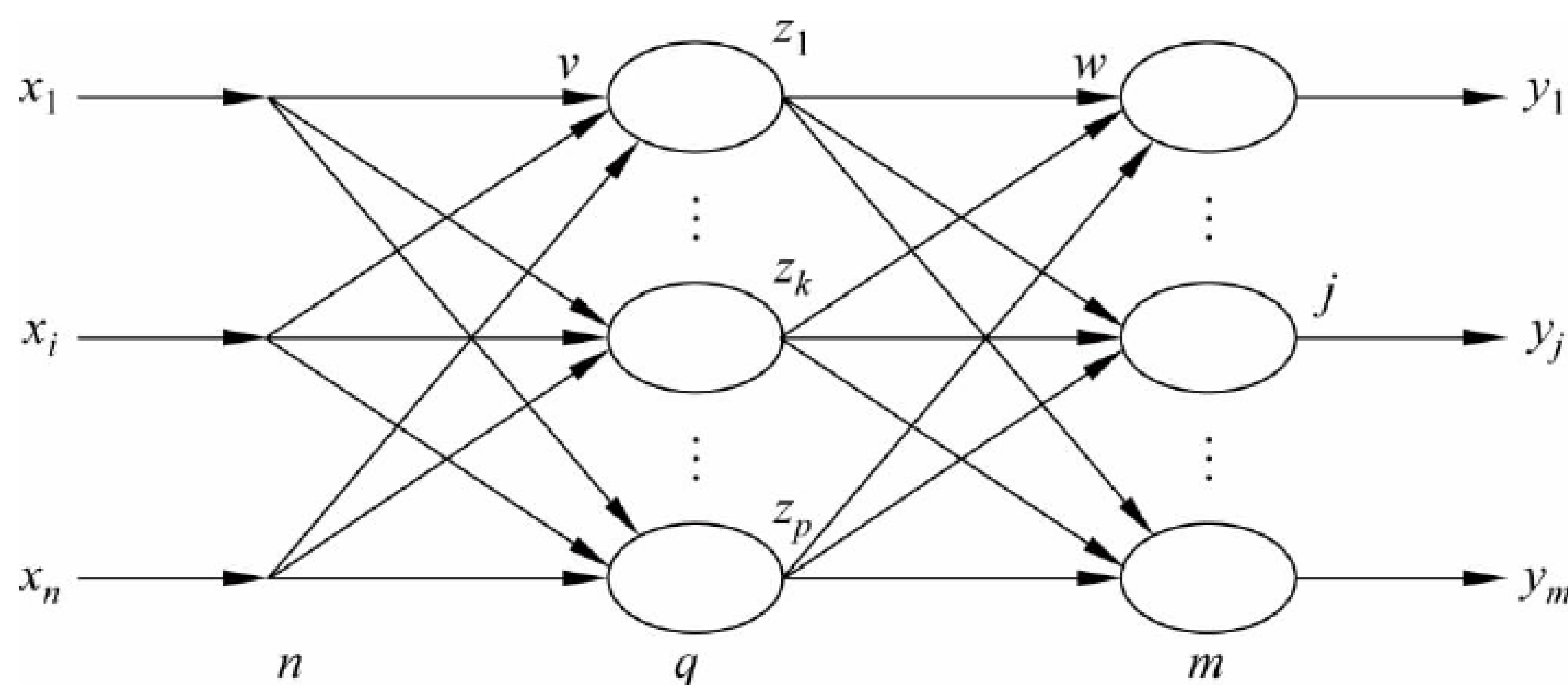


图 5-11　三层 BP 神经网络的拓扑结构示意图

网络后得到输出值为：$y_j^p(j=1,2,\cdots,m)$。采用平方型误差函数，于是得到第 p 个样本的误差值 E_p：

$$E_p = \frac{1}{2}\sum_{j=1}^{m}(t_j^p - y_j^p)^2 \tag{5-33}$$

其中：t_j^p 为期望输出。

则对于 p 个样本，整个过程中的误差可以表示为

$$E = \frac{1}{2}\sum_{p=1}^{p}\sum_{j=1}^{m}(t_j^p - y_j^p) = \sum_{p=1}^{p}E_p \tag{5-34}$$

（2）输出层权值变化

采用累积求和误差 BP 神经网络算法调整 w_{jk}，使得全局误差 E 变小，即

$$\Delta w_{jk} = -\eta\frac{\partial E_p}{\partial w_{jk}} = -\eta\frac{\partial}{\partial w_{jk}}\left(\sum_{p=1}^{p}E_p\right) = \sum_{p=1}^{p}\left(-\eta\frac{\partial E_p}{\partial w_{jk}}\right) \tag{5-35}$$

其中：η——学习效率。然后，定义误差值为

$$\delta_{yj} = \frac{\partial E_p}{\partial S_j} = -\frac{\partial E_p}{\partial y_j}\times\frac{\partial y_j}{\partial S_j} \tag{5-36}$$

式一为

$$\frac{\partial E_p}{\partial y_j} = \frac{\partial}{\partial y_j}\left[\frac{1}{2}\sum_{j=1}^{m}(t_j^p - y_j^p)^2\right] = -\sum_{j=1}^{m}(t_j^p - y_j^p) \tag{5-37}$$

式二为

$$\frac{\partial y_j}{\partial S_j} = f'_2(S_j) \tag{5-38}$$

为输出层函数的导数。

于是可以得到

$$\delta_{yj} = \sum_{j=1}^{m}(t_j^p - y_j^p)f'_2(S_j) \tag{5-39}$$

由相关定理可得

$$\frac{\partial E_p}{\partial w_{jk}} = \frac{\partial E_p}{\partial S_j} \times \frac{\partial S_j}{\partial w_{jk}} = -\delta_{yj} \times z_k = -\sum_{j=1}^{m}(t_j^p - y_j^p)f'_2(S_j) \times z_k \tag{5-40}$$

于是输出层各神经元的权值调整公式：

$$\Delta w_{jk} = \sum_{p=1}^{p}\sum_{j=1}^{m}\eta(t_j^p - y_j^p)f'_2(S_j)z_k \tag{5-41}$$

（3）隐层权值的变化

$$\Delta v_{kj} = -\eta\frac{\partial E}{\partial v_{kj}} = -\eta\frac{\partial}{\partial v_{kj}}\left(\sum_{p=1}^{p}E_P\right) = \sum_{P=1}^{P}\left(-\eta\frac{\partial E_P}{\partial v_{kj}}\right) \tag{5-42}$$

定义误差信号值为

$$\delta_{zk} = -\frac{\partial E_p}{\partial S_k} = -\frac{\partial E_p}{\partial z_k} \times \frac{\partial z_k}{\partial S_k} \tag{5-43}$$

其中式一为

$$\frac{\partial E_p}{\partial z_k} = \frac{\partial}{\partial z_k}\left[\frac{1}{2}\sum_{j=1}^{m}(t_j^p - y_j^p)^2\right] = -\sum_{j=1}^{m}(t_j^p - y_j^p)\frac{\partial y_j}{\partial z_k} \tag{5-44}$$

根据相关定理有

$$\frac{\partial y_j}{\partial z_k} = \frac{\partial y_j}{\partial S_j} \times \frac{\partial S_j}{\partial z_k} = f'_2(S_j)w_{jk} \tag{5-45}$$

式二为

$$\frac{\partial z_k}{\partial S_k} = f'_1(S_k) \tag{5-46}$$

是隐层传递函数的导数方程。

于是可以得到

$$\delta_{zk} = \sum_{j=1}^{m}(t_j^p - y_j^p)f'_2(S_j)w_{jk}f'_1(S_k) \tag{5-47}$$

由相关定理可得

$$\begin{aligned}\frac{\partial E_p}{\partial v_{kj}} &= \frac{\partial E_p}{\partial S_k} \times \frac{\partial S_k}{\partial v_{kj}} = -\delta_{zk} \times x_i \\ &= -\sum_{j=1}^{m}(t_j^p - y_j^p)f'_2(S_j)w_{jk}f'_1(S_k) \times x_i\end{aligned} \tag{5-48}$$

从而得到隐层中各神经元的权值调整公式为

$$\Delta v_{kj} = \sum_{p=1}^{p}\sum_{j=1}^{m}\eta(t_j^p - y_j^p)f'_2(S_j)w_{jk}f'_1(S_k)x_i \tag{5-49}$$

4. 神经网络的基本特性

神经网络应用于信息检索,只是该模型的一个具体应用领域。目前,对于大规模的文档集合,运用神经网络模型能否取得良好的检索性能,还有待于继续验证及相关实验数据的支持。不过,作为一类数学模型,神经网络已经在非常广泛的领域获得了惊人的成功应用。例如,手写体邮政编码判读、自动驾驶、组合优化、自动分类、生物神经活动过程模拟等。

虽然从神经生理学观点来看,神经网络模型是极端简化的,是对人脑高级神经活动的粗糙近似,但由于其对神经活动基本特征的准确捕捉,神经网络模型蕴含着巨大的理论价值与应用潜能。特别是1985年,美国加州大学的一个研究小组提出了“后向传播”(back-propagation,B-P)算法,解决了长期困扰研究人员的一个难题。B-P算法主要用于寻找一组适当的权值,以使网络具有特定的功能。B-P算法的出现,直接促成了此后有关该模型研究活动的迅猛发展。直到今天,神经网络已发展成为一个被广泛关注和探讨的、成果丰硕的研究领域。

总体上,神经网络模型的基本属性有:①非线性:人脑的思维是非线性的,故人工神经网络模拟人的思维也应是非线性的;②非局域性:非局域性是人的神经系统的一个特性,人的整体行为是非局域性的最明显体现,神经网络以大量的神经元连接模拟人脑的非局域性,它的分布存储是非局域性的一种表现;③非定常性:神经网络是模拟人脑思维运动的动力系统,它应按不同时刻的外界刺激对自己的功能进行修改,因而它是一个时变的动态系统;④非凸性:神经网络的非凸性即是指它有多个极值,也即系统具有不只一个的较稳定的平衡状态,这种属性会使系统的演化多样化,神经网络的全局优化算法就反映了这一点。

5.5 概率论检索模型

概率论模型主要基于概率论原理来理解和解决信息检索问题。在概率理论的框架基础上,目前提出的检索模型主要有早期的经典概率模型(又称为“二值独立检索模型”,即binary independence retrieval,BIR)、基于Bayesian网络的推理网络模型(inference network model)和信念网络模型(belief network model)等。

5.5.1　经典概率检索模型

1. 经典概率检索模型的基本思想

经典概率模型是一种实现简单、检索效果较好的信息检索模型，最早于 1976 年由英国城市大学的罗伯逊(S. E. Robertson)和斯帕克-琼斯(K. Sparck-Jones)提出。它是基于一个基本概率假设原理的：给定一个用户的查询请求和集合中的一篇文档 d_j，概率模型尽量评估用户找到相关的文档 d_j 的概率。模型假设相关的概率只依赖于查询请求和文档的描述。并且，假设针对查询请求 q，存在一个结果集的子集。

经典概率检索模型的基本指导思想是给定一个用户提问，则信息检索系统中存在着一个与该提问相关的理想检索命中结果集合，这里用 R 表示，如果能已知几何 R 的主要特征及其描述，则用户的检索要求便不难实现。但问题是：在用户提出检索要求时，并不知道这个理想结果几何的特性。为此，需要在检索伊始对 R 的特性进行某种猜测。根据初试的猜测，系统将检索到一个初步的命中结果集合。在此基础上，用户可以对初始检索结果集合中文档相关与否进行判断，或者由系统对检索结果文档的相关性情况进行自动判别。根据这些反馈信息，系统便可以在后续的检索处理中不断做出优化与改进，从而在多次交互之后使检索结果逐步接近该检索提问的理想命中结果集合 R。

2. 经典概率检索模型原理

(1) 原理推论一：在经典概率检索模型中，信息文档和用户检索提问仍用前述的索引词向量来表示，并且每一个索引词的权值为二值的，即 $W_{i,j}\in\{0,1\}$，$W_{i,j}\in\{0,1\}$。给定一个用户检索提问 $\boldsymbol{q}$，则相关文档集合 $\boldsymbol{d}_j$，同时令 $P(R|\boldsymbol{d}_j)$表示文档 $\boldsymbol{d}_j$ 与提问 $\boldsymbol{q}$ 相关的概率，$P(R_C|\boldsymbol{d}_j)$表示文档 $\boldsymbol{d}_j$ 与提问 $\boldsymbol{q}$ 不相关的概率，则 $\boldsymbol{d}_j$ 和 $\boldsymbol{q}$ 之间的相似度 $\mathrm{sim}(\boldsymbol{d}_j,\boldsymbol{q})$可以定义为

$$\mathrm{sim}(\boldsymbol{d}_j,\boldsymbol{q}) = P(R \mid \boldsymbol{d}_j)/P(R_C \mid \boldsymbol{d}_j) \tag{5-50}$$

(2) 原理推论二：利用贝叶斯(Bayes)公式，$\mathrm{sim}(\boldsymbol{d}_j,\boldsymbol{q})$变换为

$$\mathrm{sim}(\boldsymbol{d}_j,\boldsymbol{q}) = (P(\boldsymbol{d}_j/R)\times P(R)/P(\boldsymbol{d}_j/R_C)\times P(R)) \tag{5-51}$$

上式中，$P(\boldsymbol{d}_j/R)$表示从相关文档集合 R 中随机选择文档 $\boldsymbol{d}_j$ 的概率，或者说文档 $\boldsymbol{d}_j$ 属于相关文档集合 R 的概率；$P(\boldsymbol{d}_j/R_C)$表示从非相关文档集合 R_C 中随机选择文档 $\boldsymbol{d}_j$ 的概率，也即文档 $\boldsymbol{d}_j$ 属于非相关文档集合 R_C 的概率；$P(R)$和 $P(R_C)$则分别表示在整个文档集合随机选择一篇文档是相关和不相关时的先验概率。

(3) 原理推论三：由于 $P(R)$和 $P(R_C)$的值对于所有文档来说都是一样的，又假定索

引词之间是相互独立的，则有

$$\text{sim}(\boldsymbol{d}_j,\boldsymbol{q}) \backsim (\Pi_{g_i(d_j)=1} P(k_i \mid R)) \times (\Pi_{g_i(d_j)=0} P(\text{Nonk}_i \mid R))$$
$$/\Pi_{g_i(d_j)=1} P(k_i \mid R_C) \times (\Pi_{g_i(d_j)=0} P(\text{Nonk}_i \mid R_C)) \qquad (5\text{-}52)$$

式中，$P(k_i|R)$和 $P(\text{Nonk}_i|R)$分别表示从文档集合 R 中随机选择一篇文档，其中含有索引词 k_i 和不含有索引词 k_i 时的概率；类似地，$P(k|R_C)$和 $P(\text{Nonk}_i|R_C)$分别表示从非相关文档集合 R_C 中随机选择一篇文档，其中含有索引词 k_i 和不含有索引词 k_i 时的概率。

（4）原理推论四：考虑到有：$P(k_i|R)+P(\text{Nonk}_i|R)=1$

$$P(k_i \mid R_C) + P(\text{Nonk}_i \mid R_C) = 1 \qquad (5\text{-}53)$$

对上式取对数，再忽略掉一些常数因子，最终可得到

$$\text{sim}(\boldsymbol{d}_j,\boldsymbol{q}) \backsim \sum_{i=1}^{t} W_{iq} \times W_{ij} \times \log[(P(k_i \mid R) \times (1 - P(k_i \mid R_C)))$$
$$/(P(k_i \mid R_C) \times (1 - P(k_i \mid R))] \qquad (5\text{-}54)$$

进一步地，可以简记为

$$\text{sim}(\boldsymbol{d},\boldsymbol{q}) \backsim \sum \log[(P(k_i \mid R) \times (1 - P(k_i \mid R_C)))$$
$$/(P(k_i \mid R_C)) \times (1 - P(k_i \mid R))] \qquad (5\text{-}55)$$

（5）原理推论五：由于 R 一开始时并不是已知的，因此，要计算 $\text{sim}(\boldsymbol{d}_j,\boldsymbol{q})$，首先需要提供对概率值 $P(k_i|R)$和 $P(k_i|R_C)$的计算方法。

目前，关于 $P(k_i|R)$和 $P(k_i|R_C)$的计算方法已有多种。在开始检索前，一般做如下的简单初始假定，以启动检索进程。

① 对于所有索引词 $k_i(i=1,2,3,\cdots,t)$，$P(k_i|R)$的值都是常数，并且通常情况下规定为

$$P(k_i \mid R) = 0.5$$

② 词在非相关文档集合中的概率分布近似于索引词在全体文档集合中的概率分布，即

$$P(k_i \mid R_C) = n_i/N \qquad (5\text{-}56)$$

这里，n 和 N 的含义同前，分别表示含索引词 k_i 的文档数和系统拥有的全体文档数。

（6）原理推论六：根据上述初始假定，针对用户提问 $\boldsymbol{q}$ 的信息检索操作就可以获得一批相关文档。这里不妨用 V 来表示这批排序输出文档集合中最靠前的 r 个文档（r 是一

个预先指定的阈值)。进一步地,用 V_i 表示集合 V 中含有索引词 k_i 而形成的文档集合,V_i 中的文档数量为 r_i 个。为改善检索结果,经典概率模型需要考虑对上述 $P(k_i|R)$和 $P(k_i|R_C)$的初始计算方法加以改进。基于相关信息检索反馈调整原理,常用的改进方案主要有:

(1) $P(k_i|R)=r_i/r$

$$P(k_i \mid R_C)=(n_i-r_i)/(N-r) \tag{5-57}$$

(2) $P(k_i|R)=(r_i+0.5)/(r+1)$

$$P(k_i \mid R_c)=(n_i-r_i+0.5)/(N-r+1) \tag{5-58}$$

(3) $P(k_i|R)=(r_i+n_i/N)/(r+1)$

$$P(k_i \mid R_C)=(n_i-r_i+n_i/N)/(N-r+1) \tag{5-59}$$

采用以上任何一组 $P(k_i|R)$和 $P(k_i|R_C)$的计算公式,并多次重复检索操作及反馈调整过程,因此,概率模型系统便可有效完成各种信息检索任务。

3. 经典概率检索模型总结

从本质上来讲,信息检索是一种具有不确定性的决策判断过程。经典概率模型清楚地认识到了这种不确定性(或相关性),利用概率论原理通过赋予索引词某种概率值来表示这些词在相关信息文档集合和非相关信息文档集合中的出现概率,然后计算某一给定文档与某一给定用户提问相关的概率并做出检索决策。不同于布尔模型和向量空间模型,概率模型具有一种内在的相关反馈机制,它把检索处理过程看做是一个不断逼近并且最终确认命中信息文档集合的过程,并通过运用某种归纳式学习方法实现系统对检索结果的优化与完善。因此概率检索模型对信息检索的主要理论贡献就在于:吸收了相关反馈原理,并在理论上采用了一种更严格的决策方式。

经典概率模型虽然是一种基于贝叶斯决策的自适应模型,具有较坚实的理论基础,但就其自身来说,仍然存在着一些局限性。经典概率模型存在的局限性主要有:各种参数估计难度较大;索引词权值的计算方法为 0/1 式,没有考虑到词频等加权因素;沿用了索引词之间相互独立的基本假设。

5.5.2　贝叶斯网络检索模型

1. 贝叶斯网络检索模型概述

贝叶斯网络建立在更加完善的网络模型基础上,贝叶斯网络是人工智能领域处理不确定性问题的主要方法。不同于那些直接影响到节点诊断的方法,这种方法的节点是通过四个末端的诊断连接在一起的。正是由于这方面的优势,因此把贝叶斯网络应用于信

息检索领域是很自然的事情，为了使贝叶斯网络能够成功地应用于信息检索领域，已经取得了系列研究成果，在传统的信息检索领域先后出现了三种基于贝叶斯网络检索模型，分别是：推理网络模型(inference network model)、信念网络模型(belief network model)和贝叶斯网络检索模型(aayesian network retrieval model)。

贝叶斯网络检索模型是概率理论的一个主要研究分支。通常，Bayesian 网络可以看做是一个有向非循环图(directed acyclic graph，DAG)。图中的节点一般用来表示随机变量，有向边用于描述随机变量之间的因果关系，它由表示原因的随机变量(父节点)指向代表结果的随机变量(子节点)，而因果关系影响力的大小(或权值)则用条件概率来表示，图中没有父节点的节点称为根(root)。

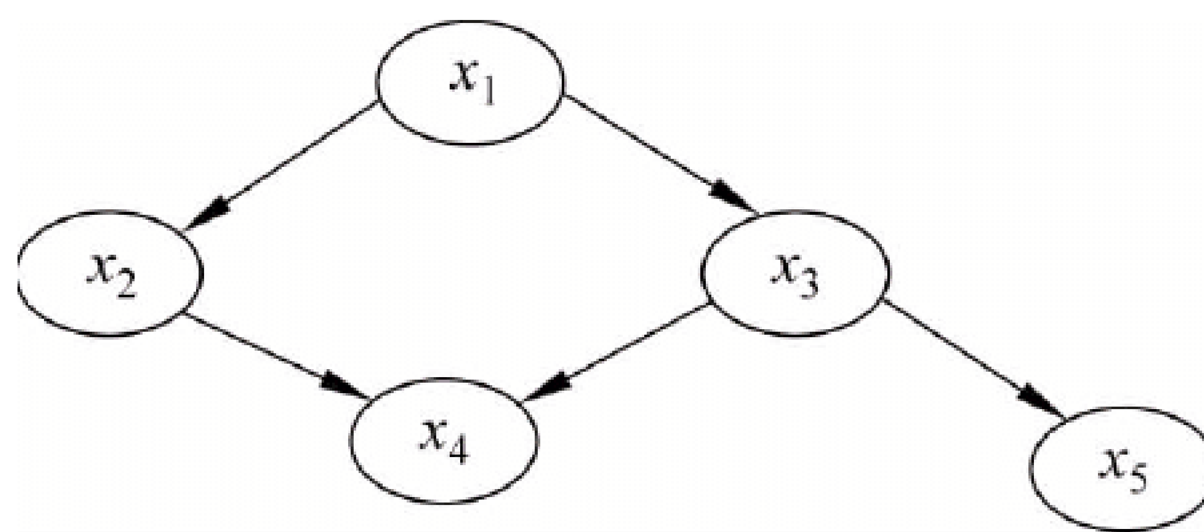


图 5-12　简单贝叶斯网络实例图

Bayesian 网络可以用联合概率分布的方式表达节点之间的依赖关系。对于图 5-12，具体表示如下：

$$P(x_1,x_2,x_3,x_4,x_5) = P(x_1)P(x_2 \mid x_1)P(x_3 \mid x_1)P(x_4 \mid x_2,x_3)P(x_5 \mid x_3) \tag{5-60}$$

上式中，$P(x_1)$称为是网络的先验概率(prior probability)，它由具体应用系统的已有知识和语义来定义或决定；其余各项则称为条件概率(conditional probability)；而联合概率分布 $P(x_1,x_2,x_3,x_4,x_5)$就描述了该 Bayesian 网络。

2. 网络检索模型推理

推理网络模型采用的是信息检索认识论的观点，该模型中文档节点用 d_j 表示，术语节点用 k_i 表示，查询节点用 q 表示。文档节点、术语节点、查询节点均与用相同符号表示的二进制随机变量相关。$U=\{k_1,k_2,\cdots,k_t\}$表示 t 维的向量空间，变量 $k_1,k_2,\cdots,k_t$ 为 U 定义了 2^t 种状态，u 表示其中一种状态。

根据查询 q 对文档 d_j 进行排序，其结果可以用来度量 d_j 的观测值为查询 q 提供了多少证据支持。在推理网络中，文档 d_j 的排序可用 $P(q|d_j)$来计算，其计算方法如下：

$$p(q \mid d_j) = \frac{P(q,d)}{P(d_j)} = \alpha P(q,d_j) \tag{5-61}$$

其中 α 是一个常数因子，因为没有对任何文档给出特定的先验概率，所以一般采用一个统一的先验概率分布，在有关推理网络的早期著作中，规定观测一篇文档 d_j 的先验概率为 $1/n$，N 为系统中的信息文档总数，因而：

$$P(d_j) = \frac{1}{N}$$

$$P(\bar{d}_j) = 1 - \frac{1}{N} \tag{5-62}$$

利用基本条件及贝叶斯定理，式(5-62)可变为下式：

$$\begin{aligned} p(q \mid d_j) &= \alpha P(q, d_j) \\ &= \beta \sum_{\forall u} P(q \mid u) \times \left(\prod_{\forall i \mid g_i(u)=1} P(k_i \mid d_j) \times \prod_{\forall i \mid g_i(u)=0} P(\bar{k}_i \mid d_j) \right) \end{aligned} \tag{5-63}$$

3. 信念网络检索模型

信念网络检索模型也是基于概率认识论描述的，但是这种模型采用的是一个明确定义的样本空间，因而产生了一种不同于推理网络的网络拓扑，即将网络中的信息文档和用户查询分离开来。

在信念网络中，术语集合 $U=\{k_1, k_2, \cdots, k_t\}$ 是一个论域(discourse)，同时为信念网络模型定义了样本空间。$u \subset U$ 是 U 的一个子集，且 $g_i(u)=1 \Leftrightarrow k_i \in u$。每个索引术语被看做是一个基本概念，因此 U 被看做是一个概念空间，概念 u 是 U 的子集。文档和用户查询用概念空间 U 中的概念表示。定义在样本空间 U 上的概率分布 P 如下所示，c 是空间 U 中的一个概念，表示一篇文档或一个用户查询：

$$p(c) = \sum_{\forall u} p(c \mid u) \times p(u) \tag{5-64}$$

$$p(u) = \left(\frac{1}{2}\right)^t \tag{5-65}$$

式(5-64)将 $p(c)$ 定义为空间 U 中 c 的覆盖度(degree of coverage)，式(5-65)表示概念空间中的所有概念均是等概率发生的。

与给定查询 q 相关的文档 d_j 的排序被理解为一种概念匹配关系，它反映了概念 q 提供给概念 d_j 的覆盖度。因此在信念网络中用 $p(d_j|q)$ 计算文档 d_j 关于查询 q 的排序。根据条件概率、公式(5-65)及贝叶斯定理可得

$$p(d_j \mid q) = \alpha P(d_j, q) = \eta \sum_{\forall u} P(d_j, u) \times P(q \mid u) \tag{5-66}$$

其中 η 为规范化因子，对概率 $p(d_j|u)$、$p(q|u)$ 的不同定义可使信念网络检索模型包括由各种经典信息检索模型(布尔模型、矢量模型、概率模型)产生的排序策略。

4. 简单贝叶斯网络检索模型

简单贝叶斯网络检索模型中的变量由两个不同的集合组成，$V=T \cup D$：集合 $T=\{T_1, T_2, \cdots, T_M\}$，集合 $D=\{D_1, \cdots, D_N\}$，T 和 D 中的变量均是二值的。变量 D_j 取值集

合为$\{\overline{d_j},d_j\}$，其中$\overline{d_j}$和d_j分别表示在给定查询下文档D_j不相关和相关。变量T_i取值集合为$\{\overline{t_j},t_j\}$，其中$\overline{t_j}$和t_j分别表示术语不相关和相关。

网络拓扑结构的建立基于以下三个假设。

(1) 如果术语T_i属于文档D_j，则术语节点T_i和文档节点D_j之间有弧。这反映了文档和其索引术语之间的依赖关系。

(2) 文档节点之间没有弧，也就是说文档节点之间的关系只是通过索引它们的术语表示出来。

(3) 已知文档D_j中索引术语是否相关的情况下，文档D_j和其他任何文档D_k是条件独立的，也就是说文档D_j是否相关只受索引它的术语影响，而不受其他文档的影响。在网络中表现为弧的指向是由术语节点指向文档节点。

由这三个假设最终确定网络的拓扑结构。网络包括两个子网：术语子网和文档子网，弧是由第一个子网中的节点指向第二个子网中的节点。该模型与推理网络模型和信念网络模型最大的区别是在网络中没有包含查询节点，也就是说该模型是独立查询的，查询只是作为证据在网络中传播。

在 BNR 模型各类节点中存储的条件概率计算如下：

(1) 对根术语节点需要存储边缘相关概率$p(t_i)$和不相关概率$p(\overline{t_i})$，可以使用$p(t_i)=(1/m)$得到$p(\overline{t_i})=1-p(t_i)=\dfrac{M-1}{M}$，其中$M$为集合中术语的数目。

(2) 对于文档节点需要估计条件概率分布$p(d_j|\pi(D_j))$，其中$\pi(D_j)$是D_j的父节点集$\Pi(D_j)$取值后的任意一种组合。因为文档节点可能有大量的父节点，所以需要估计和存储的条件概率的数目是很巨大的。因此，简单贝叶斯网络检索模型采用了专门的正则模型来表示条件概率：

$$p(d_j \mid \pi(D_j)) = \sum_{T_i \in R(\pi(D_j))} W_{ij} \tag{5-67}$$

其中$R(\pi(D_j))$是$\pi(D_j)$中相关术语的集合，权重w_{ij}满足$w_{ij}\geqslant 0$且$\sum_{T_i} W_{ij}\leqslant 1$。这样在$\pi(D_j)$中的相关术语越多，$D_j$的相关概率就越大。

简单贝叶斯网络中节点的数目通常比较大，节点之间的连接也是多路径的，每个节点也可能包含大量的父节点，所以考虑到检索的效率问题，一般的推理算法是不能使用的。因此，简单贝叶斯网络检索模型设计了特殊的推理过程可以非常有效地计算需要的概率，并且证明了得到的结果和在整个网络中实施精确推理得到的结果是一样的：

$$p(d_j \mid Q) = \sum_{T_i \in Pa(D_j)} W_{ij} \cdot p(t_i \mid Q) \tag{5-68}$$

根据术语子网的拓扑结构，则当 $T_i \in Q$ 时，$p(t_i | Q)=1$；当 $T_i \notin Q$ 时，$p(t_j | Q)=\frac{1}{M}$，这时公式(5-68)可改写为

$$p(d_j \mid Q) = \sum_{T_i \in Pa(D_j) \cap Q} W_{ij} + \frac{1}{M} \sum_{T_i \in Pa(D_j) \setminus Q} W_{ij} \tag{5-69}$$

其中权重 W_{ij} 有多种不同的计算方法。

5.6　其他检索模型的一般数学原理

集合论模型、代数模型和概率论模型的一个共同点是：都建立在对信息内容特征的标引与匹配的一般数学原理上。长期以来，对这些模型的理论探讨及实验验证，一直是信息检索领域的主要研究任务。但是，随着信息资源类型的不断丰富，信息检索的匹配机制与标准也在不断发展，除传统的信息内容特征外，信息的结构(structure)特征及其提取成为建立新型信息检索工具的另一种可供选择的匹配标准。另外，随着 WWW 网络环境的日益普及，信息检索技术也在发生着变化与调整，在 WWW 超文本技术的支持下，用户的信息检索除了通过索引文档的查询与快速匹配外，浏览方式再度兴起并流行。因此，基于信息结构特征匹配的检索模型和浏览式检索模型逐渐成为令人关注的、新的研究任务，同时基于内容的视频、音频、图像的信息检索也在快速发展，以适应多媒体信息检索的检索需要。

5.6.1　进化计算与遗传算法

进化计算(evolutionary computation，EC)这一术语是在 20 世纪 90 年代初被提出的。它是模拟生物进化过程中“优胜劣汰”的自然选择机制和遗传信息传递规律的各种算法的总称，主要用来解决实际中的复杂优化问题。目前，进化计算主要由遗传算法(genetic algorithms，GA)、遗传编程(genetic programming，GP)、进化策略(evolution strategies，ES)和进化编程(evolutionary programming，EP)等分支组成。

1. 进化计算与遗传算法的产生

生命自从在地球上诞生以来，就开始了漫长的生物进化历程，低级、简单的生物类型逐渐发展为高级、复杂的生物类型。生物进化的原因从古至今有着各种不同的解释，其中被人们广泛接受的是达尔文的自然选择学说。达尔文的自然选择学说认为：遗传和变异是决定生物进化的内在因素。其中，遗传是指父代和子代之间在性状上存在的相似现象；

变异是指父代和子代之间以及子代的个体之间，在性状上或多或少地存在差异的现象。生物的遗传特性，使生物界的物种能够保持相对的稳定；而生物的变异特性，使生物的个体产生新的性状，遗传与变异推动了生物的进化和发展。

大自然是人类获得灵感的源泉。将生物界所提供的答案应用于工程问题的求解被实践证明是一个成功的有着辉煌前景的方法。现在，人们已经认识到进化不仅仅是生命科学的范畴，进化是一种优化的过程，可以在计算机上模拟，并应用到工程领域中。早在 20 世纪 60 年代初，美国 Michigan 大学的霍兰德(J. H. Holland)教授就意识到了生物进化过程中蕴含着的朴素的优化思想，他借鉴了达尔文的生物进化论和孟德尔的遗传定律的基本思想，并将其进行提取、简化与抽象，提出了第一个进化计算算法即遗传算法。1975 年出版了他的专著 *Adaptation in Natural and Artificial Systems*，标志着遗传算法的正式诞生。在这本专著中，他称之为"Genetic Plans"，详细阐述了遗传算法的基本思想和结构框架。"Genetic Algorithms"一词首先出现在 J. D. Bagley 的博士论文中，他研究了遗传算法在博弈论(六子棋)中的参数搜索，这是遗传算法最早的应用。

图 5-13 原理性地描述了自然进化与遗传算法之间的对应关系。遗传与算法的结合体现了生物科学与计算机科学的相互渗透、相互融合。它借鉴生物的进化思想，通过计算机模拟物种繁殖过程中父代遗传基因的重新组合与"优胜劣汰"的自然选择机制的共同作用，用来解决科学与工程中的复杂问题。

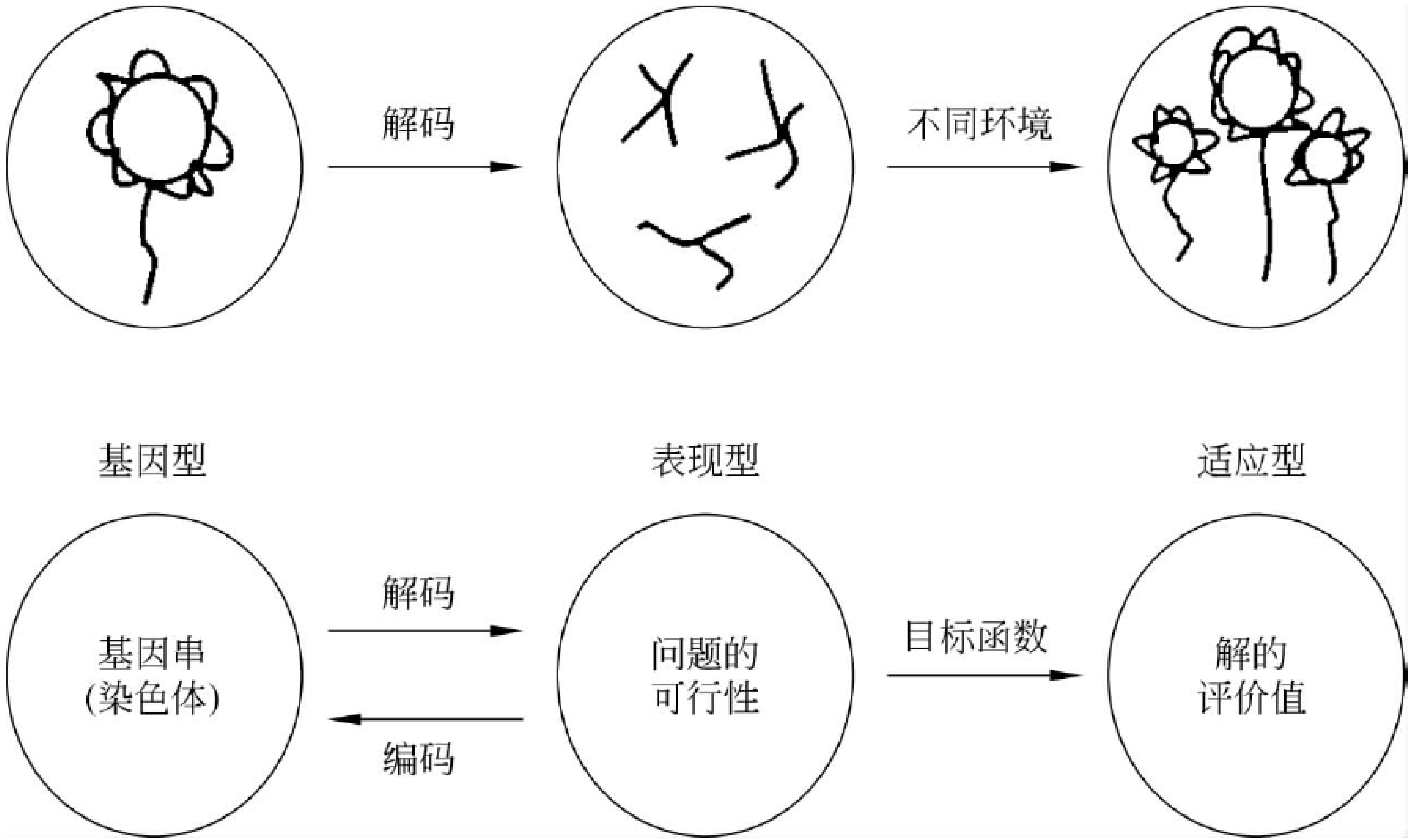


图 5-13 自然进化与遗传算法的对应关系

遗传算法产生后，在 20 世纪 80 年代以前，并没有引起人们的关注，一方面是因为它本身还不成熟；另一方面，当时的计算机容量小，计算速度慢，也使得需要较大计算量的遗传算法难以获得实际应用。但 Holland 和他的学生一直在进行坚持不懈的努力，进行了理论研究，并开拓其应用领域。直至现在，仍被认为是遗传算法理论基础的模式定理(schema theorem)就是在这个阶段提出的，它揭示了遗传算法的内部机理和解释了遗传算法的优化能力。进入 20 世纪 80 年代，遗传算法迎来了兴盛发展时期，无论是理论还是应用都成了研究热点。尤其是其应用研究显得格外活跃，给遗传算法注入了新的活力。

2. 遗传算法中的基本概念

遗传算法是遗传学和计算机科学相互结合、渗透和融合而形成的新的计算方法，其中使用了许多有关自然进化方面的基础术语，例如，

基因(gene)：控制生物性状的遗传物质的功能单位和结构单位。

染色体(chromosome)：生物遗传物质的主要载体，由多个基因组成。

基因座(locus)：染色体中基因的位置。

等位基因(alleles)：基因所取的值。

基因型(genotype)：染色体的表示模式之一，与表现型密切相关的基因组成。

表现型(phenotype)：染色体的表示模式之一，指生物个体所表现出来的性状。

同一种基因型的生物个体在不同的环境条件下可以有不同的表现型。因此，表现型是基因型与环境相互作用的结果。在遗传算法中，染色体对应的是数据或数组，在标准遗传算法中，通常是由一维的串结构数据来表现的。串上各个位置对应上述的基因座，而各位置上所取的值对应上述的等位基因。GA 处理的是染色体，或叫基因型个体，一定数量的个体组成了群体(population)，群体中个体的数量称为群体规模(population size)，而各个个体对环境的适应程度叫做“适应度”(fitness)。另外，在执行遗传算法时，必须包含两个数据转换操作。

表现型到基因型的转换：把搜索空间中的参数或解转换成遗传空间中的染色体或个体。这种转换又称编码(coding)操作，即 GA 一般不能直接处理解空间的解数据，必须通过编码将它们表示成遗传空间的基因型串结构数据。

基因型到表现型的转换：前一转换过程的逆过程，也称为译码(decoding)操作。

3. 遗传算法的基础理论

由于遗传算法是一种启发式的有向随机搜索算法，在进化过程中“是否收敛到全局最优解”成为其应用于实际问题是否成功的关键。然而，Holland 的模式定理并没有从理论上回答遗传算法的全局优化性，它只是研究了群体中部分特征模式的样本数目随进化代

数的变化规律。目前,关于遗传算法基础理论的研究在三个方面进行,即 Schema 理论的拓展与深入、遗传算法的马氏链分析、遗传算法的收敛理论。

1) Schema 理论的拓展和深入

这一方面的工作主要包括 Schema 公式的进一步讨论与拓展。Radeliffe 在其一系列工作中,把 Schema 分析进行了一般化处理,提出完整的 forma 分析理论,其主要工作集中在所谓遗传算法的欺骗函数(deceive functions)的研究上。所谓欺骗函数,就是那些对遗传算法进行误导,使其错误地收敛到非全局最优解状态的函数。一旦研究清楚一个函数是遗传算法欺骗函数的条件,也就给出了构造块假设成立的条件。研究欺骗函数问题的主要方法是 Walsh 变换。但对于确实有严重漏洞的隐含并行性原理,目前尚未有人提出改进办法,人们对这一遗传算法至关重要的优点知识加以主观信念上的默认与支持。

2) 遗传算法的马氏链分析

近年来,人们建立起了遗传算法不同形式的马氏链模型,对遗传算法的极限行为进行了各种角度的剖析。遗传算法的马氏链模型主要有三种,分别是种群马氏链模型、Vose 模型和 Cerf 扰动马氏链模型。

种群马氏链模型将遗传算法的种群迭代序列视为一个有限状态马氏链来加以研究。最早的工作属于 Goldberg,主要是运用种群马氏链转移概率矩阵的某些一般性质,分析遗传算法的极限行为,但转移概率的具体形式很难表达,这妨碍了对遗传算法有限时间行为的研究。

在 Vose 模型中,种群的状态由一个概率向量表示,概率向量的维数为所有可能个体的数目,第 i 个个体在种群的个数比例(相对概率)。当种群规模趋于无穷大时,相对概率的极限就代表了每一个个体在种群中出现的概率。无限种群规模假设下,可以导出表示种群的概率向量的迭代方程。通过对这一迭代方程的研究,可以探讨种群概率向量的迭代方程。通过对这一迭代方程的研究,可以探讨种群概率向量的不动点及其稳定性,从而导致对遗传算法极限行为的认识。虽然在无限种群假设下,Vose 模型可给出极限行为的遗传算法描述,但它们解释实际有限种群遗传算法行为的能力相对差一些。

3) 遗传算法的收敛理论

法国学者 R. Cerf 在其一系列工作中,利用 Azencott、Catoni、Trouve 等人关于模拟退火和广义模拟退火的一系列漂亮工作,将遗传算法看成一种特殊形式的广义模拟退火模型,利用动力系统的随机扰动理论,对遗传算法的极限行为及收敛速度进行了研究。尽管在 Cerf 模型中所研究的马氏链序列仍然是种群序列,但由于研究方法与种群马氏链模型的差异,我们将它称为 Cerf 扰动马氏链模型。

上述三种模型各有优缺点。种群马氏链模型最直观,因而对遗传算法行为的解释能力最强。但遗憾的是,由于对该马氏链转移概率没有一个深刻而细致的描述,目前所得结果仅仅用到了变异机制所导致的遍历性,因而只是形式上的;并且所得到的算法收敛或不收敛结果的证明方法与纯随机抽样算法相应结果的证明方法,在基本思想上这两种方法无本质区别。Vose 模型在理论上得出了一些形式复杂和漂亮的结果,但这些结果对遗传算法行为的解释性不强。Vose 模型的深入研究也许可以使遗传算法研究中用上已在群体遗传学上成功运用的随机分析方法。Cerf 的扰动马氏链模型得到了目前最完整的收敛性结果,而且有希望进一步深入。不足的是,它仍要假设种群规模趋于无穷大。

4. 生物进化思想的深层利用

虽然遗传算法已经在许多领域中获得了成功的应用,但目前仍存在几个悬而未决的问题。究其原因,主要是因为当前的遗传算法只是简单地模拟了生物的进化,对生物进化机理做了很大简化,而生物的进化是一个非常复杂的过程。

分子生物学告诉我们 DNA 的结构为由四种碱基配成的扭转阶梯螺旋,如图 5-14 所示。生物技术的发展已经使保留在化石中的 DNA 来复活生命和历史,利用 DNA 分析技术进行刑事案件分析成为可能。而遗传算法的染色体的表示则简单得多,而且用来模拟生物有性繁殖的杂交算子也多是线性串的部分交换,如图 5-15 所示。所以,要提高遗传算法的性能,必须深入地研究生物的结构与进化规律,如近年来发展起来的免疫系统模型和协同进化模型等。

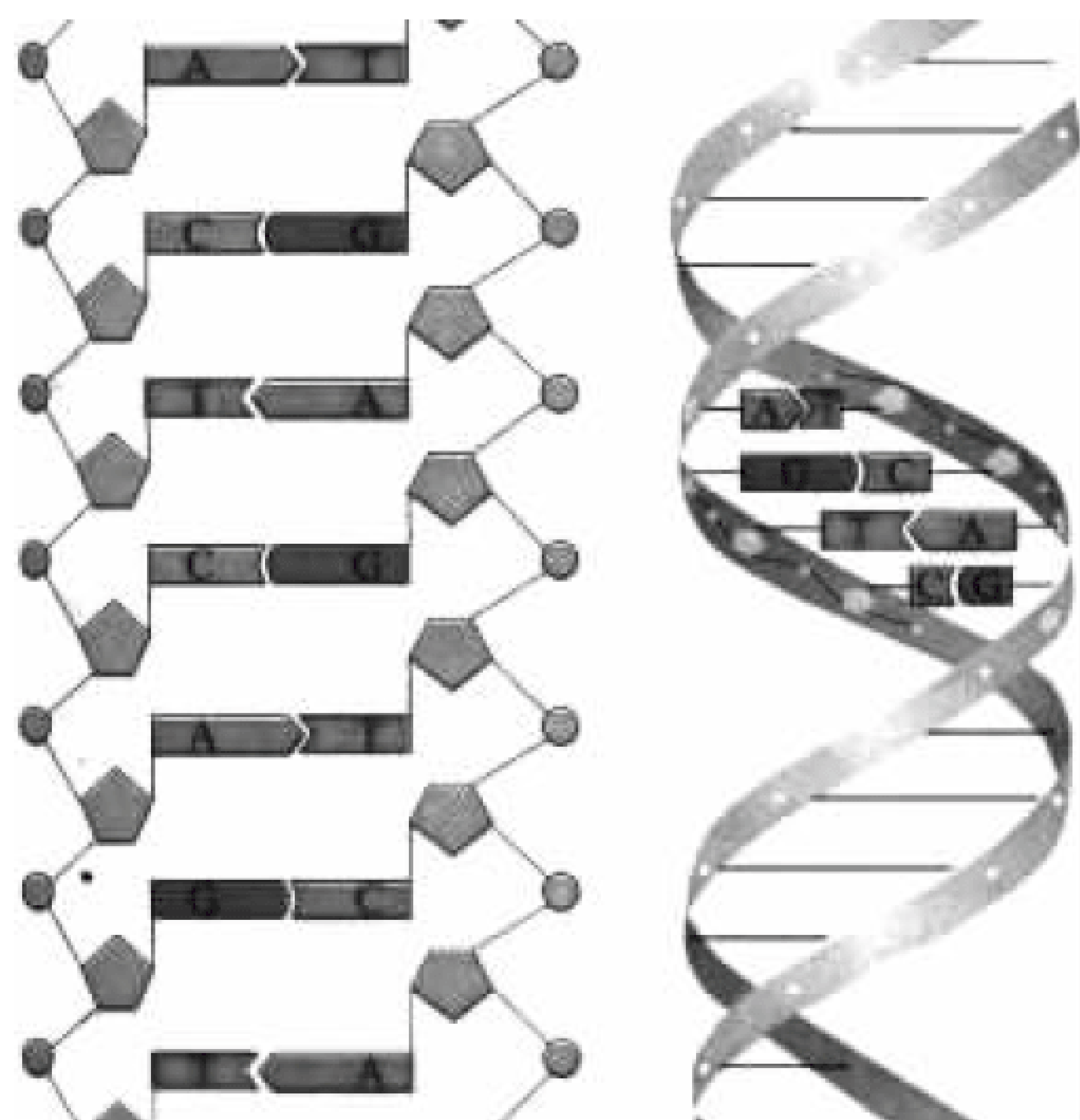


图 5-14　DNA 的双螺旋结构示意图

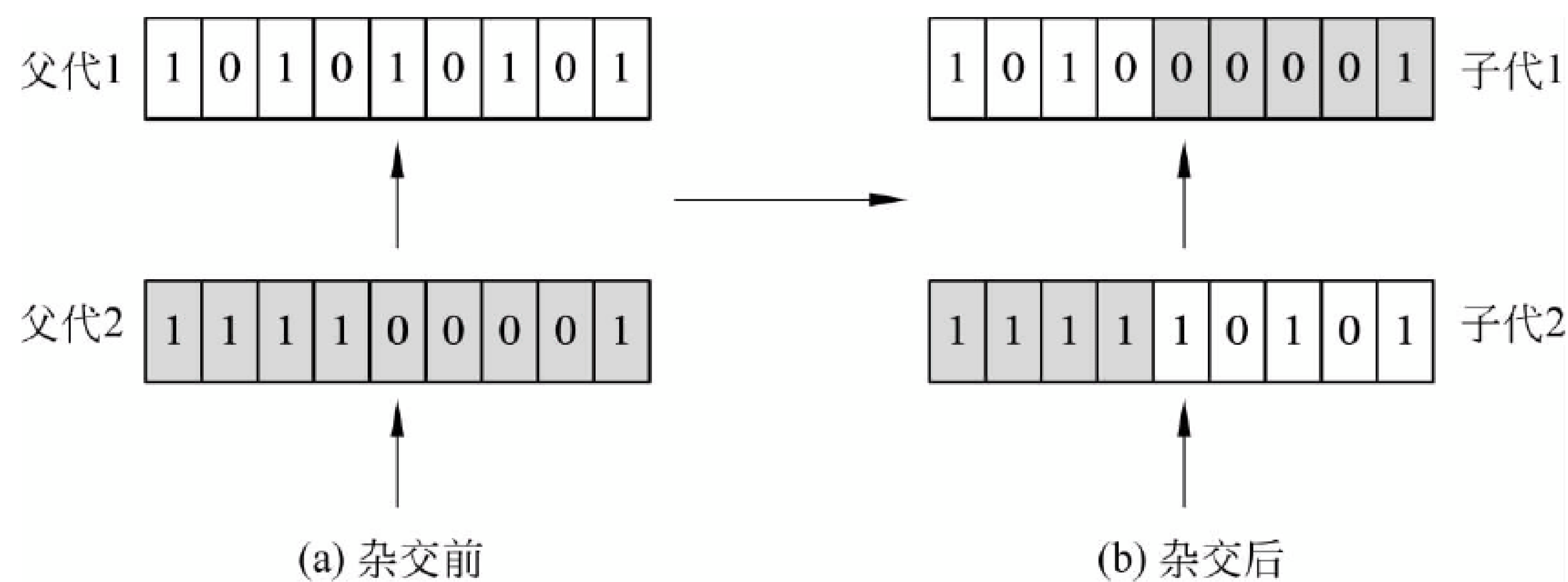


图 5-15　遗传算法的杂交算子简图

除了常用的二进制单点杂交与位变异算子，一些借鉴生物进化的新型遗传算子也已经应用在遗传算法中，如倒位（inversion ）、显性（dominance）、二倍体（diploidy）、缺失（deletion）等。

5. 遗传算法的特点与应用

遗传算法具有内在并行性(inherent parallelism)和内含并行性(implicit parallelism)。前者是指遗传算法的适应度评价是并行的，可以在并行机上进行，同时可以采用多群体进化，群体之间可以进行通信。后者是指遗传算法虽然每代仅处理 N 个个体（N 为群体规模），但却有效处理了 $O(N^3)$ 个模式。关于遗传算法的并行处理研究多集中于前者。

遗传算法的应用是一个发展最为迅速的研究方向。目前已经在模式识别、图像处理、人工智能、经济管理、机械工程、电气工程、通信、分子生物学等举不胜举的领域中获得了较成功的应用。但如何将各专业的知识融入到遗传算法的算子中，目前仍在继续研究。

概括起来说，遗传算法具有使用简单、应用范围广、鲁棒性强、易于并行化等特点。

（1）遗传算法的处理对象不是参数本身，而是对参数集进行编码的个体。这样的编码操作，使得 GA 可以直接对结构对象进行操作。所谓“结构对象”，这里泛指集合、序列、矩阵、树、图、链和表等各种一维、二维或三维结构形式的对象。

GA 的这一特点使其具有广泛的应用领域，并在组合优化问题求解、自适应控制、规划设计、机器学习和人工生命等众多领域的应用实践中，展现出了其独特的算法魅力与特色。

（2）许多传统的搜索算法都是单点搜索算法，即通过一些变动规则，问题的解从搜索空间中的当前解（点）移到另一解（点）。这种点对点的搜索算法，对于多峰分布的搜索空间常常会陷于局部的某个单峰的优解。与传统搜索算法相反，GA 是采用同时处理群体中多个个体的方法，即同时对搜索空间中的多个解进行评估。更形象地说，GA 是并行地

爬多个峰。这一特点使 GA 具有较好的全局搜索性能，可以减少陷于局部优解的风险。同时，这也使 GA 本身易于并行化。

(3) 在标准的遗传算法中，基本上不用搜索空间的知识或其他辅助信息，而仅使用适应度函数值来评估个体，并在此基础上进行遗传操作。而且，对适应度函数的唯一要求是：对于输入，可以计算出能进行比较的正值输出，即函数值≥0。GA 的这一特点使它的应用范围大大扩展。

(4) GA 不是采用确定性规则，而是采用概率的变迁规则来指导它的搜索方向，引导其搜索过程朝着搜索空间的更优化的解进行区域移动。因此，虽然看起来它是一种盲目搜索方法，但实际上却有明确的搜索方向。

下面使用一个实例“基于遗传算法思想的网络信息定题搜索应用”来说明 GA 的应用。

在 WWW 网络中，大量的网页资源通过链接形成巨大的有向图 $G=(N,E)$结构，其中，N 表示网页节点，E 表示节点之间的链接弧，并带有权值(以反映网页之间的某种关联程度)。在这样的拓扑结构中，进行定题信息搜索的目的是：在尽可能短的时间内，搜索到尽可能多的主题相关信息，同时最大限度地排除不相关信息。在搜索过程中，路径选择最为关键，并直接影响到搜索的质量和速度。

基于前面对遗传算法的理解，我们可以在定题信息搜索过程中引入遗传算法，并借助选择、交叉、变异等主要遗传算子进行搜索路径选择。算法的基本步骤设计如下。

第一，初始化。定题信息搜索的初始条件是给定待搜索主题对应的检索提问式。将检索提问式提交给某一通用搜索引擎(例如 Vista、Google 等)，搜索结果构成定题搜索的初始 URL 集合。

为提高定题搜索的效率，可对搜索引擎返回的结果进行筛选或预处理，选择一定数目的权威性较强网页的 URL 组成遗传算法需要的初始群体 $p(0)$，同时，准备好用于变异操作的网页集合 Hub。这里，Hub 页面一般是链接了多个相关主题页面、具有目录特性的页面，它们将对扩大定题搜索范围，实现全局寻优搜索具有重要作用。

第二，交叉操作。利用搜索模块下载当前群体 $p(t)$中 URL 所对应的网页，抽取网页包含的超链接，从未被搜索过的超链中挑选出被多个 $p(t)$个体页面指向的超链，组成集合 C。

第三，变异操作。按照预定的变异概率从 Hub 页面集合中提取一定数量的未被搜索的 URL，同时根据交叉概率从集合 C 中提取相应数目的 URL，共同组成新的待搜索 URL 集合 Q。

第四，选择操作。提取集合 Q 中 URL 对应页面包含的所有超链，以及超链对应的 metadata，计算各超链 URL 的适应度值，经过筛选，组成下一代群体 $p(t+1)$。这里，适应度函数 Fit 可以选择为：

$$\text{Fit}(\text{link}_i) = \text{sim}(q, \text{Metadate}(\text{link}_i)) \tag{5-70}$$

其中，Metadata(1ink)表示超链 link 的 Metadata 信息，q 为定题搜索的检索提问式，$\text{sim}(d_1, d_2)$表示 d_1 和 d_2 间的相似度。

第五，算法终止判断。算法终止参数可以有多种选择，例如进化代数 t 是否超过最大进化代数 T，已搜索网页数量是否超过用户设定的阈值，已搜索时间是否超过指定的时间值等。如果满足终止条件，则算法结束；否则，跳转到交叉操作步骤，继续进行进化过程。

上述算法设计思想充分体现了遗传算法自适应全局优化概率搜索的特点，初步的实验结果数据显示，这种定题搜索方法具有搜索范围广、查全率高等优点。

5.6.2 粗糙集理论

粗糙集(rough set，RS)理论是 20 世纪 80 年代初期由波兰数学家波拉克(Z. Pawlak)首先提出的一种数据分析理论，80 年代末期开始引起学界重视，并在数据决策与分析、模式识别、机器学习与知识发现、数据挖掘等领域得到成功应用。1995 年 ACM Communication 将粗糙集列为新出现的计算机科学研究课题；目前，该理论已成为信息科学最为活跃的一个研究领域。

1. 粗糙集理论发展概述

现实生活中有许多含糊现象并不能简单地用真、假值来表示，如何表示和处理这些现象就成为了一个研究领域——粗糙集理论。通过采取“有限的一组”和“等价关系”的系列理论，并新引入“分类”和“近似”的概念，Z. Pawlak 还扩展了以往的理论，使模糊的和不完整的数据还可以处理。早在 1904 年谓词逻辑的创始人 G. Frege 就提出了含糊一词，他把它归结到边界线上，也就是说在全域上存在一些个体既不能在其某个子集上分类，也不能在该子集的补集上分类。1965 年，Zadeh 提出了模糊集，不少理论计算机科学家和逻辑学家试图通过这一理论解决 G. Frege 的含糊概念，但模糊理论采用隶属度函数来处理模糊性，而基本的隶属度是凭经验或者由领域专家给出的，所以具有相当的主观性。20 世纪 80 年代初，波兰的 Pawlak 针对 G. Frege 的边界线区域思想提出了粗糙集(rough set)，他把那些无法确认的个体都归属于边界线区域，而这种边界线区域被定义为上近似集和下近似集的差集。由于它有确定的数学公式描述，完全由数据决定，所以更客观。

1982 年，Z. Pawlak 发表了经典论文 Rough Sets(Pawlak，1982)，标志着粗糙集理论

的诞生。由于最初的研究大多数都是以波兰文字发表的,因此该理论的研究在当时并未引起国际数学界和计算机领域的重视,研究地域仅仅局限于东欧国家。到了 20 世纪 80 年代末,粗糙集理论引起了许多数学家、逻辑学家和计算机研究人员的兴趣,他们在粗糙集理论和应用方面做了大量的研究工作。1991 年 Z. Pawlak 的专著 *Rough Set:Theoretical Aspects of Reasoning about Data* 和 1992 年 R. Slowinski 主编的关于粗糙集应用及其与相关方法比较研究的论文集的出版,对这一段时间理论和实践工作的成果做了较好的总结,推动了国际上对粗糙集理论与应用的深入研究。目前,粗糙集已经成为人工智能领域中的一个学术热点,在数据挖掘、知识获取、决策分析、过程控制等诸多领域得到了广泛的应用。我国于 2001 年 5 月在重庆召开了"第一届中国 Rough 集与软计算学术研讨会",邀请了粗糙集理论的创始人 Z. Pawlak 教授做大会报告。

2. 粗糙集理论基础

粗糙集理论是一种处理模糊和不确定信息的新的数学工具,其基本思想(Pawlak,1995)是在保持分类能力不变的前提下,通过知识的约简导出概念的分类规则。粗糙集理论最大的优点在于无须人为地额外假设条件,而是完全由已知数据来如实地回答问题,从而开辟了一条与传统智能信息处理方法所截然不同的新途径。RS 理论的基本概念主要有以下几个。

(1) 知识与知识库。一般来说,人工智能及其他复杂信息处理问题均以分类作为它们的基本机制之一。RS 理论建立在分类机制的基础上,把分类理解为等价关系,而这些等价关系将对特定问题空间进行划分。因此,在 RS 中,知识可以定义为:给定论域 U 与等价关系集合 R,R 下对数据集合 U 的划分,称为知识,记作 U/R。

可以用一个例子来说明有关的基本概念。

给定一种玩具积木的集合 $U=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8\}$,并假设这些积木有不同的颜色(红、黄、蓝)、形状(方、圆、三角)、体积(大、小)。因此,这些积木可以用颜色、形状、体积这些属性知识来描述。表 5-1 说明了这些积木的不同属性。

表 5-1　积木的总体信息集合二维表

元素集合	颜色	形状	体积	元素集合	颜色	形状	体积
x_1	红	圆	小	x_5	黄	圆	小
x_2	蓝	方	大	x_6	黄	方	小
x_3	红	三角	小	x_7	红	三角	大
x_4	蓝	三角	小	x_8	黄	三角	大

在表 5-1 中,我们定义了三个等价关系(即属性):颜色 R_1、形状 R_2 和体积 R_3,通过这些等价关系,可以得到对论域 U 形成的三个不同的划分(等价类):

$$U/R_1 = \{\{x_1, x_3, x_7\}, \{x_2, x_4\}, \{x_5, x_6\}\}$$

$$U/R_2 = \{\{x_1, x_5\}, \{x_2, x_6\}, \{x_3, x_4, x_7, x_8\}\}$$

$$U/R_3 = \{\{x_2, x_7, x_8\}, \{x_1, x_3, x_4, x_5, x_6\}\}$$

这些不同的划分构成了一个知识库,表示为

$$K = (U, R) = (U, \{R_1, R_2, R_3\}) \tag{5-71}$$

进一步地,在知识库 K 中,U/R_1、U/R_2、U/R_3 中包含的元素分别称为关于 U 的 R_1、R_2、R_3 的初等概念或初等范畴。初等范畴的交集构成基本范畴,因此,$\{x_1, x_3, x_7\} \cap \{x_3, x_4, x_7, x_8\} = \{x_3, x_7\}$表示$\{R_1, R_2\}$的基本范畴是红色三角形;而$\{x_l, x_3, x_7\} \cap \{x_3, x_4, x_7, x_8\} \cap \{x_2, x_7, x_8\} = \{x_7\}$表示$\{R_1, R_2, R_3\}$的基本范畴是红色大三角形等。

当然有些范畴在这个知识库中是无法得到的,例如,

$$\{x_1, x_5\} \cap \{x_2, x_4\} = \varnothing$$

$$\{x_1, x_3, x_7\} \cap \{x_2, x_6\} = \varnothing \tag{5-72}$$

也就是说,在我们的这个知识库中不存在蓝色圆形和红色方形的范畴。

(2) 上、下近似与粗糙集。令 X 为 U 的一个子集,R 为 U 上的一个等价关系,当 X 能表达成某些 R 基本范畴的并时,称 X 是 R 可定义的,否则称 X 是 R 不可定义的。

R 可定义集是论域的子集,它可在知识库 K 中精确地定义,而 R 不可定义集不能在这个知识库中定义。R 可定义集也称为"R 精确集",而 R 不可定义集也称为"R 非精确定义集"或"R 粗糙集"(rough set)。

对于粗糙集合可以近似地定义,我们使用两个精确集,即粗糙集的上近似(upper approximat-ion)和下近似(lower approximation)来描述。给定知识库 $K=(U,R)$,对于 U 的每一个子集 X 和一个等价关系 R,定义两个子集:

$$\begin{aligned} R-\mathrm{lower}X &= \bigcup \{Y \in U/R \mid Y \text{ 为 } X \text{ 的子集}\} \\ R-\mathrm{upper}X &= \bigcup \{Y \in U/R \mid Y \cap X = \varnothing\} \end{aligned} \tag{5-73}$$

分别称它们为 X 的 R 下近似集和 R 上近似集。

令 card 为一求集合元素个数的函数,则粗糙度可定义为

$$\alpha(X) = \mathrm{card}(R-\mathrm{lower}X)/\mathrm{card}(R-\mathrm{upper}X) \tag{5-74}$$

即将 X 对关系 R 的粗糙程度使用下近似集合元素个数与上近似集合元素个数的比值来测量。

(3) 知识约简。知识约简是粗糙集理论的核心内容之一。通常,知识库中的知识(或

属性)并不是同等重要的,甚至其中某些知识是冗余的。所谓知识约简(reduct),就是在保持知识库分类能力不变的条件下,删除其中不相关或不重要的知识。

除约简外,知识约简中还有一个基本概念——核(core),指约简时不能消去的知识特征集合,核可以作为所有约简的计算基础。

(4) 知识的依赖性。知识库中的知识之间可以是独立的或者是依赖的,而依赖程度又可能是不同的。知识 Q 是 $k(0\leqslant k\leqslant 1)$ 依赖于知识 P 的,记作 $P=>kQ$,其中：当 $k=1$ 时,称知识 Q 完全依赖于知识 P;当 $0<k<1$ 时,称 Q 粗糙(部分)依赖于 P;当 $k=0$ 时,称 Q 完全独立于 P。

(5) 知识表达系统与决策表

知识表达系统在智能数据处理中占有十分重要的地位。形式上,一个知识表达系统是一个四元组：

$$S=(U,A,V,f)$$

其中,U 为对象的非空有限集合,称为论域;A 为属性的非空有限集合;$V=\bigcup V_a(a\in A)$,V_a 真是属性 a 的值;f：$U\times A\rightarrow V$ 是信息函数,它为每个对象的每个属性赋予一个信息值。

知识表达系统的数据通常以关系表的形式表示。在各种关系表中,决策表是一类特殊而重要的知识表达系统。对于决策表而言,$A=C\cup D$,$C\cap D=\varnothing$,C 称为条件属性集,而 D 称为决策属性集。

在决策表中,最重要的是决策规则的产生。在产生决策规则之前,可首先对决策表中的属性进行约简。对决策表的处理逻辑过程一般按以下步骤进行。

① 去除重复的实例元素。

② 去除多余的属性。

③ 对每个元素删除多余的属性值。

④ 求出最小约简。

⑤ 根据最小约简,求出逻辑规则。

⑥ 区分矩阵与区分函数。

对于知识表达系统 $S=(U,A,V,f)$ 来说,其区分矩阵是一个 $n\times n$ 矩阵,其任一元素为

$$\alpha(x,y)=\{a\in A \mid f(x,a)\neq f(y,a)\} \tag{5-75}$$

因此,$a(x,y)$ 是区分对象 x 和 y 的所有属性的集合。

利用区分矩阵来表达知识有很多优点,特别是它能容易地计算约简和核。

3. 粗糙集理论的特点

作为一种研究不精确、不完整信息问题的数学工具，粗糙集理论有很多自己的特点。粗糙集理论将知识定义为不可分辨关系的族集，因此知识有了清晰的数学定义；可以很方便地用数学方法来分析处理。粗糙集理论认为知识的粒度性是造成使用已有知识不能精确地表示某些概念的原因。通过引入不可分辨关系作为粗糙集理论的基础，并在此基础上定义的上下近似等概念，粗糙集理论能够有效地逼近这些概念。不同于概率论、模糊集等其他传统数学分析工具，粗糙集理论在定量分析和处理具有不确定性和不完备性的数据时，具有非常明显的优势和特点，它通过近似集合概念来描述和表达系统的含糊性和不确定性，其表达方式更加客观，处理不确定信息的常用数据分析方法如概率论和模糊集都需要先验知识：概率论依赖于概率分布，模糊集则依赖于隶属函数，这些信息都不容易得到，而粗糙集理论对数据的分析不需要附加任何外界信息或者先验知识，所有的分析工作都能够完全基于数据对象本身完成，从而避免了主观因素的影响。正是粗糙集理论的这一独特优点，使其在数据挖掘领域迅速地脱颖而出。

但是作为一种新事物，粗糙集理论在实用中也遇到了许多困难，目前的有效途径有两条：一是粗糙集理论的进一步拓展，其次是粗糙集理论与其他方法的结合。目前基于粗糙集理论的数据挖掘主要有以下几个方面值得进一步深化。

（1）粗糙集理论和其他软计算方法进一步结合。

（2）粗糙集的基本理论中，决策信息系统的约简是 NP-Hard 问题，目前还缺乏普遍适用的算法，这是制约粗糙集理论实用化的重要方面。

（3）粗糙集理论不能直接对连续数据进行处理，必须事先对连续数据进行离散化。为了保持原有属性对决策信息系统的分辨能力，需要采用适于粗糙集的离散化算法对连续属性进行离散化。

5.6.3 浏览检索模型

一般情况下，检索方式需要通过特征提取和索引机制来实现，在用户的信息需求比较明确时，可以直接从检索系统或检索工具中进行检索和浏览，检索和浏览是用户查找和发现信息资源的两种基本手段。浏览主要依靠系统中预定义的某种信息组织和导航机制，通过用户的访问和探寻来发现一些相关的或未曾预料的有用信息。如果用户的兴趣不是提交一个对系统的查询，而是花时间浏览资源空间，以寻找所关心的文档，这种情况我们称为进行文档空间的浏览而不是搜索。因此，可以说，检索是“系统主导”方式，而浏览则是“用户主导”方式。常见的浏览模型有平坦模型、结构导向模型以及超文本模型。

在早期的计算机信息检索系统中,主要关注检索机制的建立与优化,基本上把用户从匹配过程中排除,检索处理完全由检索软件来承担,从而使得检索速度得到了极大的提高。但随着超文本技术的广泛使用,信息的浏览式查找重新引起了研究人员的注意。引入浏览机制,让用户回归并参与到信息的判断与选择过程中,在某些情形下(例如用户的信息需求不清楚或不便于表达时),可以使信息查询任务更加有效地完成。鉴于检索方式和浏览方式各有千秋,如何在系统中合理设置、平衡这两种信息查找机制,实现二者的有机结合,并在需要的时候进行自由切换,成为一个非常值得研究的信息查询问题。

1. 平坦浏览模型

该模型的思想是假设用户浏览一个具有平坦组织的文档空间。例如,文档集合可以被描述为平面(二维)上的点或是链表(一维)中的元素,用户在这些二维或一维的结构中,通过鼠标、方向键或滚动条等操作来对相关信息进行访问、阅读、浏览,以寻找有关信息。例如相关反馈过程中,用户通过在邻近文档中的浏览,查找出相关的资料或一些感兴趣的关键词。

同样,用户也可以以平面方式浏览单一的信息文档。例如,用浏览导航条浏览一个 Web 页面。

目前,这种浏览模式在信息检索系统的结果处理界面是最为流行的,但检索结果的平面式浏览,仅适用于检索结果数量较少的情形,对各种网络搜索引擎所提供的庞大检索结果集合,这样的浏览方式已成为对用户时间和精力的一个巨大浪费。平面式浏览实现方法简单,并且只能线性地按顺序进行或随机进行,效率较低。缺乏层次性的视图,容易使用户的信息浏览与查询行为迷航。

2. 结构导向浏览模型

为了对浏览的行为提供更好的支持,文档应该被组织成为如目录那样的结构。目录是类的层次结构,对文档按照主题来分类和组织。层次结构式导航是指把众多文档或信息资源组织到一个树状的类目等级体系中,用户在查找信息时可以在这样的目录结构引导下,从上到下,从宽泛到具体,逐步接近或找到所需要的有用信息。有时,对单一文档,也可以采用这样的组织方式。例如,对于一部电子图书,就可以根据其目录结构,按照章、节、小节等层次进行有关的浏览与查询导航活动。

层次结构式导航方法历史悠久,目前,在很多检索系统(例如搜索引擎)中,都设置了这样的检索初始界面。用户在查询操作伊始,即可以通过系统提供的信息资源等级分类目录,选择一个适宜的查询起点和浏览路径。为便于浏览,提高效率,通常情况下,层次式浏览还提供了一个用户访问信息的历史记录地图,以辅助用户确认浏览过的内容和次序。

层次结构式导航由于对信息集合进行了合理的分类,浏览层次与路径清晰,因而效率较高,是一种有效的信息浏览与查询机制。但是,针对大规模资源集合,如何以自动方式构建其层次组织结构,目前还是一个待解决的课题。

3. 超文本浏览模型

网状结构式浏览主要指基于超文本网页(HTML或XML)的交互性浏览模式。一般地,超文本被看做是一种由节点相互链接而形成的有向图结构。这里,节点(nodes)表示信息内容或知识单元,节点之间具有某种语义关系,用"链"(links)来表达,整个信息集合由于包含了众多信息单元而最终通过"链"形成了一个网络(network)信息架构。

超文本是一种具有巨大利用价值的信息组织与管理技术,尤其对于多媒体信息来说,更是如此。在这样的信息组织网络中,用户通过沿着不同的"链"接路径,即可探访、穿行于信息或知识的网络空间中,或浏览,或发现,或思考,在灵活地、非顺序地浏览各种相关信息的同时,还实现了与人类自身思维活动的交互作用和有机融合过程。

超文本式的导航与浏览,对于小型信息资源集合来说,无疑是一种理想的组织方式,但当集合规模较大时,超文本结构会变得非常复杂,而基于复杂的超文本结构,用户浏览时往往会出现严重的"迷路"(disorientation)现象。因此,对于超文本浏览方式,除"热键"链接技术外,一个关键问题是如何增强其导航与定位能力。目前,对导航问题提出的解决方案已有很多,例如宏观结构导航法、鱼眼视图法、浏览路标法、附加检索机制的方法等。

本章小结

信息检索技术的实现必须依靠强有力的计算机应用程序的自动执行或智能性信息处理作为支撑,而强有力的计算机应用程序必须依据数学原理及其模型方法的建立为前提。在信息检索技术中引入数学原理及其模型方法,将检索过程中的信息及其处理过程加以解释和抽象,表达成某种数学模型,再经演绎与推断,不仅能使信息检索技术作为研究对象的概念含义精确化,并且能够深刻揭示信息检索过程的显性现象与潜在的隐性规律。

布尔检索模型是一种以经典集合论和布尔代数为理论基础的非常简单的检索模型,它采用布尔代数的方法,用布尔逻辑表达式表示用户提问,通过对文献标识和提问式的比较来检索文献信息。

模糊检索数学原理是将文献看成是与提问在一定程度上相关,对于每一个标引词,都存在一个模糊的文献集合与之相关。基于模糊集合模型的检索结果是建立在文献集上的,且其隶属度就是文献集对用户提问的相关程度的模糊子集。

扩展布尔检索模型是基于布尔逻辑基本假设的一个改进模型,是一种基于布尔逻辑框架的混合布尔与向量特性的混合检索模型。扩展布尔模型是常规布尔检索精确匹配的严格性和向量处理模式提问的无结构性的折中,它用代数距离的方式来解释并放松了布尔操作的要求,因而有效融合了传统的布尔、向量等检索模型的处理思想。

检索代数模型是以线性代数、矩阵计算等数学理论为基础,利用代数论基本知识揭示信息间关系的检索模型,它在信息检索的发展中发挥着重要作用。检索代数模型主要包括向量空间模型、隐含语义索引模型、神经网络模型等具体类型。

向量空间模型是目前信息检索最常用的数学模型之一,在 WWW 信息方面,向量空间模型比布尔模型等传统模型更合适。向量空间模型(vector space model,VSM)对信息特征表达,用 TFIDF(term-fre-quency inverse-document-frequency)将 Web 页面文档转化为向量形式,再通过相关度的计算,倒排文档进行索引,从而使用户得到一个清晰的检索结果。

潜在语义索引 (latent semantic indexing,LSI)模型可以看成是一种扩展的向量空间模型,用于发现文本信息中的语义关系。潜在语义索引模式以其数学理论严谨、处理文本信息过程思路清晰得到了信息检索技术领域的重视,该方法在语言建模、视频检索等方面取得了较为成功的应用,在朴素贝叶斯分类模型、KNN 模型和 SVM 模型中都被证明是非常有效的方法。

神经网络模型主要来源于对人脑神经系统结构与功能的模拟,神经网络应用于信息检索,只是该模型的一个具体应用领域。目前,对于大规模的文档集合,运用神经网络模型能否取得良好的检索性能,还有待于验证及相关试验数据的支持。

从本质上来讲,信息检索是一种具有不确定性的决策判断过程。经典概率模型清楚地认识到了这种不确定性(或相关性),利用概率论原理,通过赋予索引词某种概率值来表示这些词在相关文档集合和非相关文档集合中的出现概率,然后计算某一给定文档与某一给定用户提问相关的概率并做出检索决策。

贝叶斯(Bayesian)网络是人工智能领域处理不确定性问题的主要方法。贝叶斯网络检索模型是概率理论的一个主要研究分支。通常,Bayesian 网络可以看做是一个有向非循环图(directed acyclic graph,DAG)。图中的节点一般用来表示随机变量,有向边用于描述随机变量之间的因果关系,它由表示原因的随机变量(父节点)指向代表结果的随机变量(子节点),而因果关系影响力的大小(或权值)则用条件概率来表示。

进化计算与遗传算法主要用来解决实际检索活动中的复杂优化问题,例如“基于遗传算法思想的网络信息定题搜索应用”。由于遗传算法是一种启发式的有向随机搜索算法,

在进化过程中是否收敛到全局最优解成为其应用于实际问题是否成功的关键。

粗糙集理论是一种新型的处理模糊和不确定信息的数学工具,其基本思想是在保持分类能力不变的前提下,通过知识的约简导出概念的分类规则。粗糙集理论最大的优点在于无须人为地额外假设条件,而是完全由已知数据来如实地回答问题,从而开辟了一条与传统智能信息处理方法所截然不同的新途径。

浏览检索模型是信息用户的一种重要信息查询与获取模型,在用户的信息需求比较明确时,可以直接从检索系统进行检索和浏览,检索和浏览是用户查找和发现信息资源的两种基本手段。浏览方式主要依靠系统中预定义的某种信息组织和导航机制,通过用户的访问和探寻来发现一些相关的或未曾预料的有用信息。

本章思考与练习题

1. 信息检索最一般的基础数学原理是什么?
2. 布尔逻辑运算符有哪三种?请分别拟定一个检索主题并用简图进行说明。
3. 布尔检索模型有何主要特点?
4. 举例说明模糊集合的信息检索应用。
5. 为什么说扩展布尔检索模型是一种混合模型?
6. 向量空间模型的含义是什么?说明基于向量空间模型的信息检索一般过程。
7. 向量空间信息检索模型有何不足之处?
8. 潜在语义索引模型的含义什么?其基本思想是什么?
9. 与传统的向量空间模型相比,LSI有哪些优点?
10. 神经网络模型的基本思想是什么?如何理解反向传播学习算法(BP)的含义?
11. 神经网络模型有哪些基本属性?
12. 说明经典概率检索模型的基本指导思想。
13. 贝叶斯网络检索模型的基本含义是什么?
14. 贝叶斯网络检索模型的网络拓扑结构的建立基于哪些假设?
15. 进化计算与遗传算法对信息检索技术研究有何作用?
16. 粗糙集理论的基本思想是什么?有哪些主要特点?

第6章　文本分类与文本索引构建

文本分类(text categorization,TC)又称为文本自动分类,它是信息检索和文本挖掘的重要基础。分类任务就是通过学习得到一个目标函数,即分类模型,通过此分类模型把每个属性集映射到一个预先定义的类中。文本分类是在预定义的分类体系下,根据文本的特征即文本的内容,将给定文本与一个或多个类别文本进行相互关联的过程。文本自动分类能较好地解决大量检索文档信息归类的问题并可以应用到很多方面,如文本信息组织、文本识别、智能搜索、邮件过滤等,因此对文本分类的学习与研究具有重要的理论意义和实用价值。文本分类是一种具有指导性的自动学习机制,是根据一个已经被标注的训练文档集合找到其文档特征和文档类别之间的关系模型,然后利用这种学习到的关系模型对未被标注的文档进行类别判断。文本分类作为文本信息过滤、文本信息检索、文本数据创建、数字化图书馆建设、大型专用数据库检索系统或网络搜索引擎构建等领域的技术基础,有着广泛的应用前景。

文本分类技术属于一种有监督(supervised)机器学习方法。一般来说,文本分类的过程如下:获取训练文本集,训练文本集由一组经过预处理的文本特征向量组成,每个训练文本样本有一个类别标号。利用训练文本集对初始分类模型进行训练并得出分类判别模型。利用训练得到的分类判别模型对其他待分类文本进行自动分类和归类。由于文本分类的主要任务就是在预先给定的类别体系下,通过对有标记文本集的学习,将文本集中未标记的文本对象映射到预设的类别中,因此文本分类能很好地满足信息检索对文本信息组织提出的重要需求。

文本分类的方法有决策树分类方法、k-最邻近分类方法、KNN 算法和朴素贝叶斯分类方法等。不同方法的精度各不相同,适用的领域也不一样。在这些方法中,朴素贝叶斯分类方法的验证结果比预设想象要好,其目标是在测试数据或新数据(new data)上获得高精确率的结果。文本分类的关键问题是如何构造一个分类函数或分类模型(也称为分类器),并利用此分类模型将未知文本映射到给定的类别空间。分类器的构造方法有多种,主要有统计方法、机器学习方法、神经网络方法等。

朴素贝叶斯分类器是贝叶斯分类器中最常用的方法,也是一种基于概率统计的方法。

朴素贝叶斯分类方法是基于条件“独立性假设”,因此它适合于处理属性个数较多的分类任务,而文本分类正是适应了这种多属性的分类任务,因此朴素贝叶斯成为文本分类的一种常用分类方法,它也是目前公认的一种简单有效的概率分类方法。

6.1 文本分类概述

分类是指将给定对象归入一个或者多个类别的过程,通常来说,类别往往是一个一般的主题领域,而不是很狭窄的固定范围,面向文本的分类任务则称为文本分类。分类不一定要使用计算机,很多分类任务都是通过人工来完成的,但是人工分类的方法一旦要规模化则开销会很大。可以采用直接利用固定查询将其想象成某种规则来进行分类,这些规则一般是由人工编写的。这些规则通过关键词的某种组合来代表一个类别。人工编写的规则具有很好的扩展性,但是创建和长时间维护这些规则需要很高的人力成本。

除了效率低的手工分类和人工编写规则之外,还存在高效率的基于机器学习的分类方法。当学习方法基于统计时,这种方法也称为统计文本分类。在统计文本分类中,对于每个类别需要一些良好的文档样例。由于需要人来标注训练文档,所以对人工分类的需求仍然存在。标注是指对每篇文档赋予类别标签的工作。文本分类任务从数学的角度来看就是一个映射过程,可以使用如下的数学模型来描述。

文本分类中,给定文档 $d \in X$ 和一个固定的类别集合 $C=\{c_1, c_2, \cdots, c_j\}$,其中 X 表示文档空间,类别(class)也通常称为 category 或 label。一般文档空间 X 是某种类型的高维空间,而类别通常由人们根据具体应用需求来定义,比如 China 类及有关 computer hardware 的文档类。给定已经标识好类别的训练集(training set) $D=\langle d, c\rangle$,其中 $\langle d, c\rangle \in X \times C$,例如

$$\langle d, c\rangle = \langle \text{Beijing joins the World Trade Organization}, \text{China}\rangle$$

表示一句话文档 Beijing joins the World Trade Organization 被标记为 China 类。利用某种学习方法(learning method)或学习算法(learning algorithm)可得到某个分类函数(classification function) γ,γ 可以将文档映射到类别:

$$\gamma: X \rightarrow C$$

由于监督者(定义类别体系并标注训练集的人)在学习过程中起到类似导师的作用,所以这种类型的学习称为有监督的学习。这里把有监督学习方法记为 Γ,故有 $\Gamma(D)=\gamma$。Γ 以训练文档集 D 为输入,返回学习到的分类函数 γ。

下面简述分类任务:给定文档集合 $D=\{D_1, D_2, \cdots, D_n\}$,$D_i$ 表示第 i 篇文档。D 由 n

篇文档组成；预先定义的文档类别集合 $C=\{C_1,C_2,\cdots,C_{|C|}\}$。假设文档集合与类别存在一个未知的目标函数：

$$\Phi: D\times C\rightarrow\{\text{True},\text{False}\} \tag{6-1}$$

文本分类任务可以描述为要努力找到的一个函数：

$$\hat{\Phi}: D\times C\rightarrow\{\text{True},\text{False}\} \tag{6-2}$$

使$\hat{\Phi}$尽量逼近未知的目标函数 Φ，$\hat{\Phi}$称为分类器（classifier）或者模型（model）。如果 $\Phi(D_i,C_j)=\text{True}$，则文档 D_i 属于类别 C_j；$\Phi(D_i,C_j)=\text{False}$，则文档 D_i 不属于类别 C_j。也就是说，文本分类的最终目的就是要找到一个有效的隐射函数，准确地实现 $D\times C$ 到值 True 或 False 的映射。

中文文本不像英文文本那样单词与单词之间有空格，因此中文文本分类需要进行中文分词。如今，中文分词的技术已趋于成熟，主要有中国科学院计算技术研究所研制的汉语词法分析系统 ICTCLAS。结合中文文本的特点，逐步形成了中文文本信息的分类研究体系。一个完整的中文文本分类系统通常由几个紧密联系的功能模块组成。

（1）文本预处理：文本预处理是对文档进行分词，去除停用词，其中中文分词是文本预处理的首要步骤。

（2）文本表示：文本表示是文本分类的基础。要将计算机技术应用到文本分类上，必须把文档转化为计算机容易处理的表示形式。目前使用最普遍的文本表示方式是向量空间模型。

（3）文本特征选择：特征选择的目的是为了维数约简，从文档中抽取出若干最有利于文本分类的特征项。

（4）特征权重计算：特征权重是用于衡量某个特征项在文档表示中的重要程度或者区分能力的强弱。

（5）分类器学习训练：分类器学习训练的目的是建立分类器，是文本分类的核心问题。利用一定的学习算法对训练样本集进行统计学习，估算出分类器的各个参数，从而建立对训练集进行学习训练的自动分类器。

（6）测试与评价：利用学习训练阶段建立的分类器，对测试集文档进行分类测试。在完成训练和测试后，选择合适的评价指标对分类器的性能进行评价。如果分类性能不符合要求，需要返回前面步骤。

按照文本分类的工作顺序，文本分类可以分为三大阶段。

第一阶段：将文本表示成文本向量。这个阶段需要完成的工作是先对文本进行预处

理,然后进行特征选择和特征权重计算后,将文本转换成向量空间模型的形式。

第二阶段:学习训练阶段。选择分类方法,使用已经表示成文本向量的训练集来建立分类模型。

第三阶段:测试与评价。将第二阶段建立好的分类模型运用于测试集来检验分类效果,并使用评价指标对分类模型的性能进行评价。

目前基于统计机器学习的文本分类技术相对成熟,被广泛应用于很多检索系统和网络检索工具。其中包括基于概率方法的朴素贝叶斯分类器、基于实例的 k 近邻分类器、基于统计学习理论和结构风险最小原理基础上的支持向量机方法。还有其他的分类方法,包括线性分类器、回归模型、神经网络、决策树方法等。基于机器的学习方法很少考虑文本语义信息,目前研究者大多是把语义分析、概念网络和机器学习方法相结合,从概念级来获取文本的语义,进而提高文本分类的效果。

6.2 朴素贝叶斯文本分类

朴素贝叶斯文本分类(naive Bayes classification,NBC)的一个前提假设是:在给定的文档集中,文档属性是相互独立的。朴素贝叶斯分类是建立在经典的贝叶斯概率理论基础之上,其基本思想是利用特征项和类别的条件概率来估算给定文档的类别概率,是一种基于概率统计的分类方法。朴素贝叶斯分类是贝叶斯学习方法中最常用的方法,也是一种简单而又非常有效的分类方法。贝叶斯分类模型是一种典型的基于统计方法的分类模型。贝叶斯定理是贝叶斯理论中最重要的一个公式,是贝叶斯学习方法的理论基础,它将事件的先验概率与后验概率巧妙地联系起来,充分利用先验信息和样本数据信息确定事件的后验概率。

6.2.1 贝叶斯分类器

朴素贝叶斯文本分类的主要工作是设计分类器,目前贝叶斯分类器主要有两种。

一种是朴素贝叶斯分类器,它是贝叶斯分类模型中最简单、最有效而且在实际使用中非常成功的分类器,其性能可以与神经网络、决策树相媲美。朴素贝叶斯分类模型基于假定特征向量的各分量间相对于决策变量是相对独立的,即条件独立性假设。尽管这一假定在一定程度上限制了朴素贝叶斯分类模型的适用范围,但在实际应用中,降低了贝叶斯网络构建的复杂性。朴素贝叶斯分类模型已成功地应用到聚类、分类等数据挖掘、大数据处理的查询与搜索任务中。

为了突破朴素贝叶斯分类器的独立性假设条件的限制，人们通过改变其结构假设的方式来达到目的。例如半朴素贝叶斯分类器 SNBC(semi-naive bayesian classifier)、树扩张型 TAN(tree-augmented bayesian classifier)及增强型贝叶斯分类器 BAN(Bayesian network augmented naive Bayes)等。这些分类器具有如下特点。

(1)贝叶斯分类并不把一个对象绝对地指派给某一类，而是通过计算得出属于某一类的概率，具有最大概率的类便是该对象所属的类。

(2)一般情况下在贝叶斯分类中所有的属性都潜在地起作用，即并不是一个或几个属性决定分类，而是所有的属性都参与分类。

(3)贝叶斯分类对象的属性可以是离散的、连续的，也可以是混合的。

另一种是贝叶斯网络分类器，贝叶斯网络又称为信念网络，它是基于后验概念的贝叶斯定理。贝叶斯网络是一个有向无环图，其中节点代表论域中的变量，有向弧代表变量的关系，变量之间的关系强弱由节点与其父节点之间的条件概率来表示，通过贝叶斯网络可以准确地反映实际应用中变量之间的依赖关系。贝叶斯网络可用于分类、聚类、数据挖掘、大数据处理、人工神经网络、预测和因果关系分析等。贝叶斯网络分类器具有很强的学习、推理能力，能很好地利用先验知识。

6.2.2 条件概率和乘法定理

在事件 A 已经发生的条件下事件 B 发生的概率，称为事件 B 在给定事件 A 的条件概率(也称为后验概率)，记作 $P(B|A)$。相应地，$P(A)$称为无条件概率(也称先验概率)，条件概率可以依照下式进行计算：

$$P(B \mid A) = \frac{P(AB)}{P(A)} \tag{6-3}$$

由条件概率可求得概率的乘法定理：

$$P(A \cdot B) = P(B \mid A)P(A) \tag{6-4}$$

对于 n 个事件 $A_1, A_2, \cdots, A_n$，$n \geqslant 2$，则有

$$P(A_1, A_2, \cdots, A_n) = P(A_n \mid A_1 \cdot A_2 \cdots A_{n-1})P(A_{n-1} \mid A_1 \cdot A_2 \cdots A_{n-2}) \cdots P(A_2 \mid A_1)P(A_1) \tag{6-5}$$

6.2.3 极大后验假设和极大似然假设

定义：极大后验假设：在许多学习场景中，学习器考虑候选假设集合 H，并在其中寻找给定的数据 D 时的可能性最大假设 $h \in H$。这样具有最大可能性的假设被称为极大后

验假设(maximum a posteriori,MAP),记作:h_{MAP}。

$$h_{\text{MAP}} = \arg\max_{h \in H} p(h \mid D) = \arg\max_{h \in H} \frac{p(D \mid h)p(h)}{p(D)}$$
$$= \arg\max_{h \in H} p(D \mid h)p(h) \tag{6-6}$$

去掉 $P(D)$,因为它不依赖于 h 常量,上式就是一个原始的分类模型,贝叶斯分类就是根据上述 MAP 假设找出的新实例最有可能的分类。所有对贝叶斯分类模型的研究工作都是以此假设为前提条件的。在某些情况下,可假定 H 中的每个假设都有相同的先验概率(即对 H 中的任意的 h_i 和 h_j,有 $P(h_i)=P(h_j)$),这时可以把式(6-4)进一步进行简化,只考虑 $p(h|D)$来寻找极大可能假设,$p(D|h)$常被称为给定 h 时数据 D 的似然度,而使得 $p(D|h)$最大的假设成为极大似然假设,记作:h_{ML}。

$$h_{\text{ML}} = \arg\max_{h \in H} p(D \mid h) \tag{6-7}$$

与机器学习问题相联系,把数据 D 称为某目标函数的训练样本,把 H 称为候选目标函数空间。

6.2.4 贝叶斯定理

定义 1:如果 P 是 R 上的一个实值函数,即对每一个 $A \in R$,有一个实函数 $P(A)$与之对应,并且满足以下三点。

非负性:对任意 $A \in R$,$P(A) \geqslant 0$。

规范性:$P(R)=1$。

可加性:若 $A_1, A_2, \cdots, A_n, \cdots$是 R 中的两两不相容的事件,则

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) \tag{6-8}$$

称 P 是(Ω,R)上的一个概率(测度),$P(A)$称为事件 A 的概率,三元组(Ω,R,P)称为概率空间。

定义 2:设(Ω,R,P)为一概率空间,A、$B \in R$,且 $P(A) \geqslant 20$,则

$$P(B \mid A) = \frac{P(AB)}{P(A)} \tag{6-9}$$

称为已知 A 发生时 B 的条件概率。

全概率公式:设 $A_1, A_2, \cdots, A_n \in R$,两两不相容,$P(A) \geqslant 0$,$i=1,2,\cdots,n$,且 $\bigcup_{i=1}^{n} A_i = \Omega$,则对任何事件 $B \in R$,有

$$P(B)=\sum_{i=1}^{n}P(B\mid A_i)P(A_i) \tag{6-10}$$

贝叶斯公式：设 $A_1,A_2,\cdots,A_n\in R$，两两不相容，$P(A)\geqslant 0$，$i=1,2,\cdots,n$，则对于任何满足 $P(B)\geqslant 0$ 的 B，$B\in R$，有

$$P(A_j\mid B)=\frac{P(B\mid A_j)P(A_i)}{\sum_{i=1}^{n}P(B\mid A_i)P(A_i)} \tag{6-11}$$

6.2.5　多项式朴素贝叶斯

此方法中，文档 d 属于类别 c 的概率的计算方法如下：

$$P(c\mid d)\propto P(c)\prod_{1\leqslant k\leqslant n_d}P(t_k\mid c) \tag{6-12}$$

其中，$P(t_k|c)$是 t_k 出现在类 c 文档中的条件概率，也可以把 $P(t_k|c)$视为当正确类为 c 时 t_k 的贡献程度。$P(c)$是文档出现在类 c 中的先验概率。如果根据文档的词项并不能清晰地区分它属于哪一类时，我们就选择先验概率最大的那个类。$\langle t_1,t_2,\cdots,t_{nd}\rangle$是 d 中的词条，它们是分类所用词汇表的一部分，n_d 是 d 中所有词条的数目。例如，对于单句文档 Beijing and Taipei join the WTO，如果将 and 和 the 视为停用词过滤掉，那么这里的 $\langle t_1,t_2,\cdots,t_{nd}\rangle$就可以是〈Beijing，Taipei，join，WTO〉，其中 $n_d=4$。

在文本分类中，我们的目标是找出文档最可能属于的类别。对于 NB 分类来说，最可能的类是具有 MAP(maximum a posteriori，最大后验概率)估计值的结果 C_{map}：

$$c_{\text{map}}=\arg\max_{c\in C}\hat{P}(c\mid d)=\arg\max_{c\in C}\hat{P}(c)\prod_{1\leqslant k\leqslant n_d}\hat{P}(t_k\mid c) \tag{6-13}$$

由于我们不知道参数的真实值，所以上述公式中采用了从训练集中得到的估计值来代替 P。

对所有的 $1\leqslant k\leqslant n_d$，计算其对应的条件概率的乘积，这可能会导致浮点数下界溢出。因此，更好的方法是引入对数，从而将原公式的计算转变成多个概率的对数之和。由于 $\log(xy)=\log(x)+\log(y)$，且 log 是单调递增函数，因此具有较高概率对数值的类别就是最有可能的类别。因此，大多数 NB 在实现时所求的最大值实际是

$$c_{\text{map}}=\arg\max_{c\in C}\hat{P}(c\mid d)=\arg\max_{c\in C}\left[\log\hat{P}(c)+\sum_{1\leqslant k\leqslant n_d}\log P(\hat{t_k}\mid c)\right] \tag{6-14}$$

对于上式，有个简单的解释，条件参数 $\log\hat{p}(t_k|c)$表示的是 t_k 在类别 c 中的权重，而对数先验值 $\log\hat{p}(c)$表示的是类别 c 的相对频率的一个权重值。相对于低频类而言，高频

类更可能是正确类。类别的对数先验值和词项在类别中权重累加求和之后就得到了文档属于类别的可能程度，式(6-14)选择最可能的类别作为最终的类别。

我们先基于这种直观解释来使用上述公式，这实际上是多项式 NB 模型的一个解释。如何估计参数$\hat{P}(c)$及$\hat{P}(t_k|c)$呢？首先我们使用最大似然估计(MLE)，它实际上最后算出的是相对频率值，这些值能使训练数据的出现概率最大。MLE 估计下的类别先验概率为

$$\hat{P}(c)=\frac{N_c}{N} \tag{6-15}$$

其中，N_c 是训练集合中 c 类所包含的文档数目，而 N 是训练集合中的文档总数。条件概率$\hat{P}(t|c)$ 的估计值为 t 在 c 类文档中出现的相对频率：

$$\hat{P}(t \mid c)=\frac{T_{ct}}{\sum_{t'\in V}T'_{ct'}} \tag{6-16}$$

其中，T_{ct}是 t 在训练集合 c 类文档中出现的次数，在对每篇文档计算时用的是其在文档中多次出现的词频。这里我们引入了位置独立性假设(positional independence assumption)，在该假设下，T_{ct}是 t 在训练集某类文档中所有位置 k 上的出现次数之和。这样对于不同位置上的概率值都采用相同的估计办法，比如，如果某词在一篇文档中出现过两次，分别在 k_1 和 k_2 的位置上，那么假定$\hat{P}(t_{k_1}|c)=\hat{P}(t_{k_2}|c)$。

最大似然估计(MLE)的一个问题是：对没有在训练集中出现的词项和类别项来说，其 MLE 估计值为 0。比如，如果在训练集上，WTO 仅仅在 China 类文档中出现，那么对于其他类(如 UK)，采用 MLE 估计的概率值就会为 0，即$\hat{P}(\mathrm{WTO}|\mathrm{UK})=0$。

现在，假定有一篇单句文档为 Britain is a member of the WTO，那么按照公式(6-12)来计算其属于 UK 类的条件概率值就为 0。很显然，由于文档中包含 Britain，此时应该为其属于 UK 类的条件概率赋予一个较高的值。也就是说，此时不能对 WTO 属于 UK 类的概率值赋 0，因为一旦出现 0 值，其他词项的概率再高也没有意义。出现零概率的主要原因来自数据的稀疏性(sparseness)，即训练集合永远都不可能大到所有罕见事件都能出现，这样就难以计算这些事件的频率。比如，上面要计算的 WTO 出现在 UK 类文档中的频率。

为了去掉零概率，一个简单的方法是采用加一平滑(add-one smoothing)或拉普拉斯平滑(Laplace smoothing)，即在每个数字上加 1：

$$\hat{P}(t \mid c) = \frac{T_{ct} + 1}{\sum_{t' \in V}(T_{ct'} + 1)} = \frac{T_{ct} + 1}{\sum_{t' \in V} T_{ct'} + B} \tag{6-17}$$

其中，$B=|V|$是词汇表中所有词项的数目。加一平滑可以认为是采用均匀分布作为先验分布（每个词项在每个类中出现一次），然后根据训练数据进行更新得到的结果。

到此，已给出了文本训练和应用贝叶斯分类器的所有环节，完整的算法描述如图 6-1 所示。

```
TRAINMULTIONOMIALNB(C, D)
1. V ← EXTRACTVOCABULARY(D)
2. N ← COUNTDOCS(D)
3. for each c ∈ C
4. do N_c ← COUNTDOCSINCLASS(D, c)
5. prior[c] ← N_c/N
6. text_c ← CONCATENATETEXTOFALLDOCSINCLASS(D, c)
7. for each t ∈ V
8. do T_ct ← COUNTTOKENSOFTERM(text_c, t)
9. for each t ∈ V
10. do condprob[t][c] ←
11. return V, prior, condprob
APPLYMULTINOMIALNB(C, V, prior, condprob, d)
1. W ← EXTRACTTOKENSFREMDOC(V, d)
2. for each c ∈ C
3. do score[c] ← log prior[c]
4. for each t ∈ W
5. do score[c] += log condprob[t][c]
6. return arg max_{c∈C} score[c]
```

图 6-1　多项式贝叶斯训练和应用分类算法

6.3　朴素贝叶斯分类模型改进

朴素贝叶斯分类器是基于一个简单的假定：在给定分类特征条件下属性值之间是相互条件独立的。在现实世界中，它的属性独立性假设使其无法表示实际应用中各属性之间的依赖关系，影响了它的分类性能。因此需要针对实际应用对朴素贝叶斯分类器模型进行改进，使之在属性独立性假设不满足的情况下依然具有较高的分类精度。

6.3.1　改进方法

朴素贝叶斯分类器的本质是一种具有很强限制条件的贝叶斯网络分类器，但是它限

制条件太强，不适于很多现实应用情况。然而完全无限制条件的贝叶斯网络也是不现实的，因为学习这样的网络非常耗时，其时间复杂度为属性变量的指数级，并且空间复杂度也非常高。因此，研究朴素贝叶斯分类器的改进模型，只能从这两者之间来考察，即研究其有较宽松条件限制的贝叶斯网络分类器。

（1）属性删除方法：适用于存在冗余属性的情况。学者 Langley 和 Sage 提出了一种基于属性删除的选择性贝叶斯分类器。当存在一些属性依赖于其他属性，特别是存在冗余属性时，属性删除方法确实能够改善朴素贝叶斯分类器的预测精度。

（2）构造新属性或概率调整方法：适用于某些属性依赖于其他属性时。学者 Pazzani 等提出了通过相互依赖的属性构造一个新属性，并用新属性取代原来相互依赖的那些属性方法。这种方法也可视为事先的条件概率调整方法。学者 Wang 和 Webb 等提出了一种半懒惰式（semi-lazy）的限制性贝叶斯网络分类器的条件概率调整方法，在某些情况下可以减小误分类率。

（3）局部朴素贝叶斯分类器：适用于属性之间相互依赖情形比较复杂的情况。这种方法是为属性变量的每一种取值（或某个范围）建立一个朴素贝叶斯分类器，也就是说，单一的全局朴素贝叶斯分类器被许多局部朴素贝叶斯分类器所代替，将属性独立性假设放宽到只要局部属性独立就可以了。学者 Kohavi 将朴素贝叶斯分类器和决策树相结合，用一棵决策树来分割实例空间，在每个叶子节点上建立局部朴素贝叶斯分类器，学者 Zheng 等利用懒惰式学习策略提出了一种懒惰式贝叶斯规则（lazy bayesian rule）学习方法，该方法将懒惰式方法应用到局部朴素贝叶斯规则的归纳中，该算法虽然较大地提高了分类精确度，但是效率很低。

（4）树扩张型贝叶斯方法：学者 Friedman 等提出了一种树扩张型贝叶斯方法。这种方法的基本思路是放宽朴素贝叶斯的独立性假设条件，扩展朴素贝叶斯的结构，使其能够容纳属性间存在具有某种特征的依赖关系。利用条件相互信息（conditional mutual information）建立属性之间的依赖关系矩阵，构造一棵最大权生成树作为一个分类器。由于限制每个属性节点最多有一个非类变量（类标识）的父节点，也就是说每个属性节点最多仅依赖于一个非类标识节点，使其表示依赖关系的能力受到限制。

（5）限定性双层贝叶斯分类模型：学者石洪波等提出了一种限定性的双层贝叶斯分类模型，这种方法的出发点是通过对属性空间的搜索，找出一些对其他属性有较强影响的属性，那么所有其他的属性仅通过与这些属性的关联就可以将重要的依赖关系表示出来。

6.3.2 朴素贝叶斯分类的提升模型

对朴素贝叶斯分类模型进行“提升”(boosting)是在不改变独立性假设的前提下提高分类性能的一种方法。提升方法的主要思想是从训练实例中学习一系列的分类器。每一个分类器根据前一个分类器错误分类的实例,对训练实例的权重进行修正,再学习新的分类器。例如,学习得到分类器 KH 后,增加了由 KH 导致分类错误的训练实例的权值,并且通过重新对训练实例计算权值,再学习下一个分类器 KH+1。这个过程重复 T 次,从这个系列的分类器中可以综合得出最终的分类器。

提升算法实现了对分类问题的处理,算法描述如下。

Begin

Input：N 个训练实例：$D=\{(x^1,c^1),\cdots,(x^N,c^N)\}$以及待分类实例,由于包括 N 个训练实例上的分布 D：w,w 为训练实例的权向量。

T：训练重复次数(或轮数)

$$\text{Output：} h(x) = \arg\max \sum_{t=1}^{T} \left(\log \frac{1}{\beta^{(t)}}\right) I(h^{(t)}(x) = c) \tag{6-18}$$

其中 $I(\omega)$是实例函数,当 $\omega=T$ 时 $I(\omega)=1$,否则 $I(\omega)=0$。

步骤：

初始化训练实例的权向量,$W_t=1/N, i\in(1,\cdots,N)$

For $t=1$ to T

给定权值 $W_i^{(t)}$ 得到一个假设 $H^{(t)}:X\rightarrow C$ 估计假设 $H^{(t)}$的总体误差：

$$e^{(t)} = \sum_{i=1}^{N} w_i^{(t)} I(c^i \neq h^{(t)}(x^i)) \tag{6-19}$$

则计算 $\beta^{(t)}=e^{(t)}/(1-e(t))$,

然后计算下一轮样本的权值：

$$w_i^{(t+1)} = w_i^{(t)} (\beta^{(t)})^{1-I(C^{I=h^{(t)}}(x^{(i)}))} \tag{6-20}$$

规范化 $w_i^{(t+1)}$,使其总和为 1

End for

假设每一个分类器都是有用的,则 $e^{(t)}<0.5$。也就是说,在每一次分类的结果中,正确分类的样本个数始终大于错误分类的样本个数。可以看出,此时 $\beta^{(t)}<1$,那么当对某个训练实例 x^i 分类结果不正确时,实例函数 $I(\omega)=0$,导致 $w_i^{(t+1)}$ 增加,因此满足了提升的思想。上述提升朴素贝叶斯分类器的时间复杂度是 0(Tnf),其中 f 是每个样本的属性的

个数。在一般情况下,提升后的分类性能有了较大的提高,但是这种提升方法也存在不足:一是不能捕捉属性间的相关性,也就是说没有突破条件独立性假设的限制;二是当训练集中存在噪音数据时,提升方法会把噪音数据当成有用的信息通过权值而放大,从而降低提升的性能。

6.3.3 基于特征相关的改进加权朴素贝叶斯分类

朴素贝叶斯文本分类方法是基于特征项间独立的假设,但是这与实际情况不一定相符,为此研究出一种加权朴素贝叶斯算法,对后验概率计算中的每个条件概率项进行加权,并且对不同的特征项提供不同的加权值,从而使得特征项之间是不独立的,它们对类别的重要程度是不一样的。基于特征相关的改进加权朴素贝叶斯算法,在传统"词频-逆文档频率"(TF-IDF)权重的基础上,考虑到类内和类间分布,同时根据特征项之间的相关程度,对它们的权重进行调整,突出相关性比较大的特征项权重,从而提高了加权朴素贝叶斯的分类能力。

加权朴素贝叶斯文本分类。朴素贝叶斯分类方法认为所有条件属性对决策属性的分类重要性是一致的(权重均为1),这种方式使得冗余的、与分类无关的、相互影响的以及被噪声污染的特征和其他特征具有相同的地位,并使得分类的正确性降低,实际上,有些因素对分类影响大一些,而另外的要小一些。基于此提出将各种特征加权算法与朴素贝叶斯分类器相结合,对不同的特征根据其分类重要性赋予不同的权值,使朴素贝叶斯扩展为加权朴素贝叶斯以提高分类器的性能。加权朴素贝叶斯模型大多为

$$C(x) = \arg\max P(C_i) \prod_{\substack{k=1 \\ C_i \in C}}^{k=n} P(x_k \mid c_i) W_{j=k} \tag{6-21}$$

其中 $W_{j=k}$ 是特征项 t_j 在类别 c_k 中的权重,权重越大,该特征项对分类的影响越大。

特征权重的计算方式有很多种,比如布尔权重、词频权重、TF-IDF 权重等。而 TF-IDF 权重应用最广泛,因为它将词频和逆文档频率结合使用,克服了其他权重计算的缺点,TF-IDF 计算的归一化公式如:

$$W_i = \frac{\mathrm{TF}(t_i) \times \mathrm{IDF}(t_i)}{\sqrt{\sum_{i=1}^{n} (\mathrm{TF}(t_i) \times \mathrm{IDF}(t_i))^2}} \tag{6-22}$$

$$\mathrm{IDF}(t_i) = \log\left(\frac{N}{n_i} + L\right)$$

其中 $\mathrm{TF}(t_i)$ 是特征项 t_i 的词频,$\mathrm{IDF}(t_i)$ 是逆文档频率,在公式中,L 的取值通过实验来确定。N 为文档集的总文档数,n_i 为出现特征项 t_i 的文档数。IDF 算法的核心思想

是，在大多数文档中都出现的特征项不如只在一小部分文档中出现的特征项重要。IDF 算法能够弱化一些在大多数文档中都出现的高频特征项的重要程度，同时增强一些在小部分文档中出现的低频特征项的重要程度。

6.4　贝努利文本分类模型

另外一种文本分类模型方法是多元贝努利模型(multivariate Bernoulli model)或者直接称为贝努利模型(Bernoulli model)。它等价于二值独立模型，对于词汇表中的每个词项都对应一个二值变量，1 和 0 分别表示词项在文档中出现和不出现。图 6-2 给出了基于贝努利模型的 NB 分类器的训练和测试算法。贝努利模型和多项式模型具有一样的时间复杂度。

```
TRAINBERNOULLINB(C, D)
1. V ← EXTRACTVOCABULARY(D)
2. N ← COUNTDOCS(D)
3. for each c∈C
4. do N_c ← COUNTDOCSINCLASS(D, c)
5.     prior[c] ← N_c/N
6.     for each t∈V
7.     do N_ct← COUNTDOCSINCLASSCONTAININGTERM(D, c, t)
8.     condprob[t][c] ← (N_ct+1) (N_c+ 2)
9. return V, prior, condprob
APPLYBERNOULLINB(C, V, prior, conprob, d)
1. V_d ← EXTRACTTERMSFROMDOC(V, d)
2. for each c∈C
3. do score[c] ← log prior[c]
4.     for each t∈V
5.     do if t∈V_d
6.     then score[c]+=log condprob[t][c]
7.     else score[c]+=log(1－condprob[t][c])
8. return argmax_{c∈C} score[c]
```

图 6-2　基于贝努利模型的 NB 算法的训练及分类过程

不同的生成模型也意味着不同的参数估计策略和分类规则。贝努利模型中 $\hat{P}(t|c)$ 利用类 c 文档中包含 t 的文档数的比率来计算。而与之形成鲜明对比的是，多项式模型中计算的是 t 出现的次数占类 c 文档中所有词条数目的比率。当对测试文档进行分类时，贝努利模型只考虑词项的出现或不出现(即二值)，并不考虑出现的次数，而多项式模型中则要考虑出现次数。这样做的结果是，当对长文档进行分类时，采用贝努利模型往往会犯很多错误。比如，可能会因为 China 在文档中一次出现而将整本书归于 China 类。两种模型

(多项式模型与贝努利模型)对于未出现词项在分类中的使用也不相同。未出现的词项在多项式模型中并不影响分类效果,但是在贝努利模型中计算 $P(c|d)$ 时要以一个因子来参与计算,其主要原因是,贝努利模型对词项的未出现也要显式建模。

例如,对于表 6-1 中的例子采用贝努利模型进行计算,对于先验概率,我们同多项式模型中一样估计,即 $\hat{P}(c)=3/4$, $\hat{P}=1/4$。

表 6-1 用于参数估算的文本词项数据

文本集	文档 ID	文档中的词项	C=China 类?
训练集	1	Chinese Beijing Chinese	Yes
	2	Chinese Chinese Shanghai	Yes
	3	Chinese Macao	Yes
	4	Tokyo Japan Chinese	no
测试集	5	Chinese Chinese Chinese Tokyo Japan	?

条件概率为

$\hat{P}(\text{Chinese} \mid c) = (3+1)/(3+2) = 4/5$;

$\hat{P}(\text{Japan} \mid c) = \hat{P}(\text{Tokyo} \mid c) = (0+1)/(3+2) = 1/5$;

$\hat{P}(\text{Beijing} \mid c) = \hat{P}(\text{Macao} \mid c) = \hat{P}(\text{Shanghai} \mid c) = (1+1)/(3+2) = 2/5$;

$\hat{P}(\text{Chinese} \mid \bar{c}) = (1+1)/(1+2) = 2/3$;

$\hat{P}(\text{Japan} \mid \bar{c}) = \hat{P}(\text{Tokyo} \mid \bar{c}) = (1+1)/(1+2) = 2/3$;

$\hat{P}(\text{Beijing} \mid \bar{c}) = \hat{P}(\text{Macao} \mid \bar{c}) = \hat{P}(\text{Shanghai} \mid \bar{c}) = (0+1)/(1+2) = 1/3$。

这个问题中有三篇文档词项属于 c 类,1 篇文档属于非 c 类,另外由于对每个词项都只考虑出现与不出现两种情形,因此公式(6-17)中的常数 B 为 2。因此,测试文档分别属于两个类别的得分为

$$
\begin{aligned}
\hat{P}(c \mid d_5) \propto\ & \hat{P}(c) \cdot \hat{P}(\text{Chinese} \mid c) \cdot \hat{P}(\text{Japan} \mid c) \cdot \hat{P}(\text{Tokyo} \mid c) \\
& \cdot (1-\hat{P}(\text{Beijing} \mid c)) \cdot (1-\hat{P}(\text{Shanghai} \mid c)) \cdot (1-\hat{P}(\text{Macao} \mid c)) \\
=\ & 3/4 \cdot 4/5 \cdot 1/5 \cdot 1/5 \cdot (1-2/5) \cdot (1-2/5)(1-2/5) \\
\approx\ & 0.005
\end{aligned}
$$

类似地有

$$\hat{P}(\bar{c} \mid d_5) \propto 1/4 \cdot 2/3 \cdot 2/3 \cdot 2/3 \cdot (1-1/3) \cdot (1-1/3)(1-1/3) \approx 0.022$$

因此，根据上述结果，分类器最终会将测试文档归为非 c 类。当只关注词项出现与否而不考虑词项频率时，Japan 和 Tokyo 对于 $\bar{c}$ 来说是正向标志特征（$2/3>1/5$），而 Chinese 属于 c 类和非 c 类的条件概率的差异还不足以影响分类的结果。

6.5　多项式文本分类模型与贝努利文本分类模型的性质比较

多项式模型是：$P(d|c)=P(\langle t_1,\cdots,t_k,\cdots,t_{n_d}\rangle|c)$，贝努利模型为：$P(d|c)=P(\langle e_1,\cdots,e_i,\cdots,e_M\rangle|c)$，其中$\langle t_1,\cdots,t_{nd}\rangle$是在 d 中出现的词项序列，$\langle e_1,\cdots,e_i,\cdots,e_M\rangle$是一个 M 维的布尔向量，表示每个词项在文档 d 中存在与否。

解决文本分类问题的一个关键步骤是选择文档的表示方法，而$\langle t_1,\cdots,t_{nd}\rangle$和$\langle e_1,\cdots,e_i,\cdots,e_M\rangle$正是两种不同的文档表示方法。在第一种表示方法中，文档空间 X 是所有词项序列的集合，也可以说是所有词条序列的集合。为了减少参数的数目，下面引入朴素贝叶斯的条件独立性假设，即给定类别时，假设属性值之间是相互独立的：

$$\text{多项式模型}\quad P(d \mid c) = P(\langle t_1,\cdots,t_{n_d}\rangle \mid c) = \prod_{1\leqslant k\leqslant n_d} P(X_k = t_k \mid c) \tag{6-23}$$

$$\text{贝努力模型}\quad P(d \mid c) = P(\langle e_1,\cdots,e_M\rangle \mid c) = \prod_{1\leqslant i\leqslant M} P(U_i = e_i \mid c) \tag{6-24}$$

上式中引入了两类随机变量 X_k 和 U_i，这样在两个不同的文本生成中，模型就更清晰。X_k 是文档在位置 k 上的随机变量，$P(X_K=t|c)$表示的是一篇 c 类文档中词项 t 出现在位置 k 上的概率。随机变量 U_i 对应词项 i，当词项在文档中不出现时取 0，出现时取 1。$P(U_i=1|c)$表示的是 t_i 出现在 c 类文档中的概率，这时可以是在任意位置上出现多次。

例如图 6-3 与图 6-4 所示，对五个词项属性（对应多项式模型）和六个二值属性（对应贝努利模型），China 类对应都有一个概率值。一篇 China 类文档中包含 Taipei 的事实并不会增加或者减少该文档包含 Beijing 的可能性。

在检索文档分类实践当中，文本数据上的条件独立假设并不成立，词项之间存在条件依赖。但是可以看到，尽管采用了条件独立性假设，NB 模型也表现出很好的性能。即使是采用条件独立性假设，但假如在文档中每个位置 k 上的概率分布不同，则对于多项式模型来说仍然具有太多的参数需要估计。词项在文档中的出现位置本身并不包含任何对分

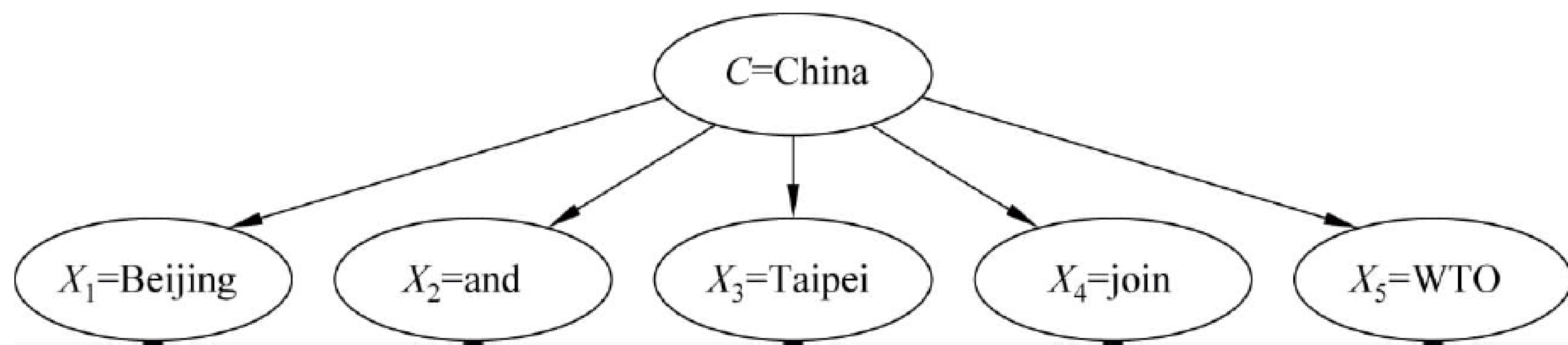


图 6-3　多项式贝叶斯文本分类实例

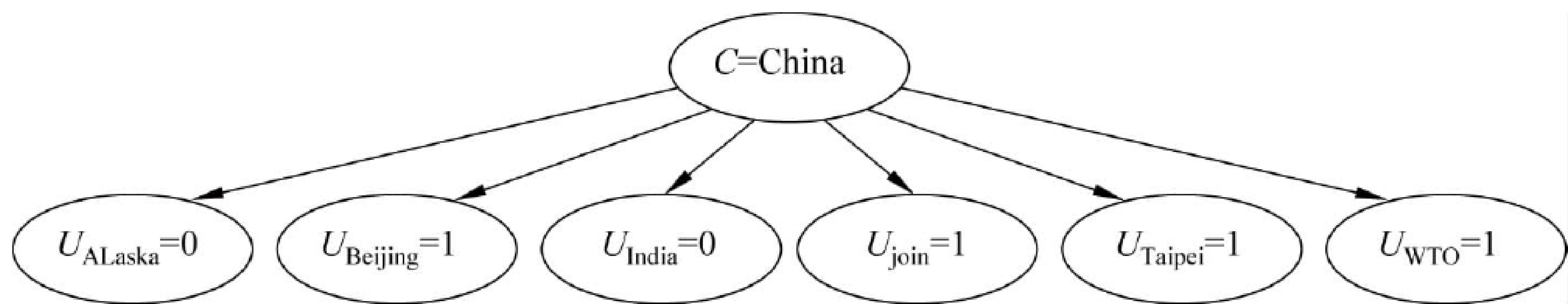


图 6-4　贝努利文本分类实例

类有用的信息。尽管 China sues Japan 和 Japan sues China 不同，但是 China 是在文档中第一个位置还是在第三个位置出现对于多项式分类来说毫无分别，这是因为多项式分类中对每个词项都是独立看待的，条件独立性假设对上述处理方法提供了有效支持。

另一方面，如果假设在不同位置 k 上词项分布不同的话，那么就要估计每个 k 的一系列参数。比如，bean 出现在 coffee 类文档的第一个位置和出现在其第二个位置的概率是不同的，其他位置可以依次类推，这会再次导致数据估计中的稀疏性问题。

在多项式模型中，首先以概率 $P(c)$ 来选择一个类别 $C=c$，其中 C 是一个随机变量，然后根据模型生成一篇文档。接着，对于文档的 n_d 个位置，在每个位置 k 上以概率 $P(X_k=t_k|c)$ 生成词项 t_k。并且对于给定的 c，每个 X_k 的分布是一样的。在图 6-3 所示的例子中，给出了单句文档 Beijing and Taipei join WTO 的生成过程，其中 $\langle t_1, t_2, t_3, t_4, t_5\rangle = \langle$Beijing, and, Taipei, join, WTO$\rangle$。

对于一个完全确定的文档生成模型而言，还需要对 $P(n_d|c)$ 这个长度分布进行定义。如果没有这个分布，那么该多项式分布就是一个词条的生成模型而不是一个文档的生成模型。

在贝努利模型（如图 6-4 所示）文档的生成过程中，首先以概率 $P(c)$ 来选择一个类别 $C=c$，然后对词典中的每个词项 $t_i(1\leqslant i\leqslant M)$ 都产生一个对应的二值变量 e_i。在图 6-3 的例子中，仍然以单句文档 Beijing and Taipei join WTO 为例，说明了 $\langle e_1, e_2, e_3, e_4, e_5, e_6\rangle = \langle 0,1,0,1,1,1\rangle$ 的生成过程（其 and 被看成停用词）。下面用表 6-2 说明两个模型之间的

比较结果，其中包括计算公式和决策规则的比较。

表 6-2　多项式模型和贝努利模型的比较

比较项	多项式模型	贝努利模型
事件模型	词条生成模型	文档生成模型
随机变量	$X=t$，当且仅当 t 出现在给定位置	$U_t=1$，当且仅当 t 出现在文档中
文档表示	$d=\langle t_1,\cdots,t_k,\cdots,t_{nd}\rangle,t_k\in V$	$d=\langle e_1,\cdots,e_i,\cdots,e_M\rangle,e_i\in\{0,1\}$
参数估计	$\hat{P}(X=t\|c)$	$\hat{P}(U_i=e\|c)$
决策规则：最大化	$\hat{P}(c)\prod_{1\leqslant k\leqslant n_d}\hat{P}(X=t_k\mid c)$	$\hat{P}(c)=\prod_{t_i\in V}\hat{P}(U_i=e_i\mid c)$
词项多次出现	考虑	不考虑
文档长度	能够处理更长文档	最好处理短文档
特征数目	能够处理更多特征	特征数目较少效果更好
词项 the 的估计	$\hat{P}(X=\text{the}\|c)\approx 0.05$	$\hat{P}(U_{\text{the}}\|c)\approx 1.0$

6.6　文本分类特征选择

6.6.1　文本分类特征选择的作用

在文本分类中，特征项应该具有如下特性：特征项要能明确标识文本信息内容；特征项具有将目标文本与其他文本进行区分的能力；特征项的个数不能太多，即维度不能太多，否则会耗费大量的计算资源；特征项分离要比较容易实现。

如果把文档信息中所有的词都作为特征项，那么特征向量的维数将过于巨大，从而导致计算量巨大，在这样的情况下，要完成文本信息自动分类几乎是不可能的。特征选择的任务就是在不改变文本核心内容信息的前提下，尽可能地减少要处理的特征项数量，从而降低向量空间维数，进行简化计算以提高文本处理的速度和效率。朴素贝叶斯分类模型是建立在属性之间条件独立性假设之上，因此特征选择的好坏与否对分类精度有较大影响。

特征选择(feature selection)是从训练集合出现的词项中选出一部分子集的过程。在文本分类过程也仅仅使用这个子集作为特征。特征选择是文本信息模式识别的关键问题之一，特征选择结果的好坏直接影响着分类器的分类精度和泛化性能。下面首先分析特征选择方法的框架，然后从信息检索和搜索策略与评价准则两个角度对特征选择方法进

行分析。

特征选择有两个主要目的：第一，通过减少文本内容有效的词汇空间来提高分类器训练和应用的效率，这对于除NB之外的训练开销较大的分类器来说尤为重要。第二，特征选择能够去除噪音数据特征，从而提高分类的精度。

6.6.2 特征选择的方法

特征选择的基本框架见图6-5。

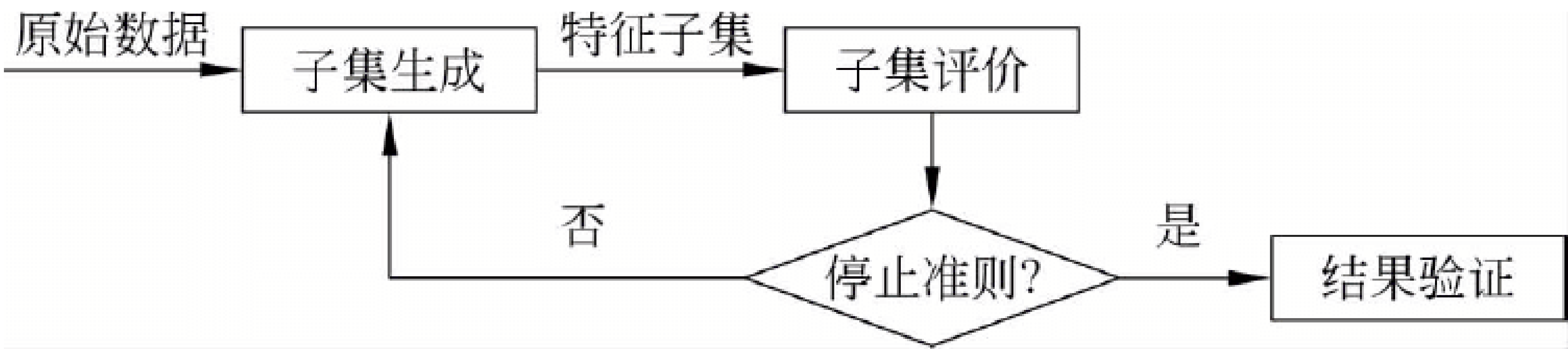


图6-5 特征选择的基本框架

从特征选择的基本框架可以看出，特征选择方法有四个基本步骤：候选特征子集的生成（自动搜索策略）、评价准则、停止准则和验证方法。经典特征选择定义为从 N 个特征集合中选出 M 个特征的子集，并满足条件 $M \leqslant N$。它包括特征提取和特征选择两个方面：特征提取广义上指的是一种变换，将处于高维空间的样本通过映射或变换的方式转换到低维空间，达到降维的目的；特征选择指从一组特征中去除冗余或不相关的特征来降维。二者常结合使用，如先通过变换将高维特征空间映射到低维特征空间，然后再去除冗余的和不相关的特征来进一步降低维数。

特征选择主要用于排除确定的特征空间中那些被认为无关的或是关联性不大的特性，于是经常会使用特征独立性假设以简化特征选择，以达到计算时间和提高计算质量的折中目的。因此，目前在对文本特征空间所采取的特征选择算法一般是构造一个评价函数，对特征集中的每个特征进行独立的评估。这样每个特征都获得一个评估分，然后对所有的特征按照其评估分的大小进行排序，选取预定数目的最佳特征作为结果的特征子集。所以，选取多少个最佳特性以及采用什么评价函数，都需要针对某一个具体的问题通过试验来决定。

对于基本的特征选择算法，简单地说，给定类别 c，对词汇表中的每个词项 t，我们计算效用指标 $A(t,c)$，然后从中选择 k 个具有最高值的词项作为最后的特征，其他的词项则在分类中都被忽略。特征选择算法有三种不同的效用指标：互信息 $A(t,c)=I(U_t,C_c)$、χ^2 统计量 $A(t,c)=\chi^2(t,c)$ 及词项频率 $A(t,c)=N(t,c)$。

6.6.3　特征选择方法类型

特征选择需要解决两个问题：一是确定选择算法，在允许的时间内，以可以忍受的代价找出最小的、最能描述类别的特征组合；二是确定评价标准，衡量特征组合是否最优，得到特征获取操作的停止条件。因此，一般分两步进行特征获取：先产生特征子集，然后对子集进行评价，如果满足停止条件，则操作完毕，否则重复前述两步直到条件满足为止。

第一种，按照特征子集的形成方式，特征获取方法可分为穷举法、启发法和随机法三类。启发式方法为一种近似算法，具有很强的主观倾向。实际应用中通过采用期望的人工机器调度规则，重复迭代产生递增的特征子集。特征个数为 N 时，复杂度一般小于或者等于 $O(2^N)$。这种方法实现过程比较简单而且快速，在实际中应用非常广泛，如向前（向后）选择、决策树法、Relief 方法及其改进方法等。但是它不能保证结果最优，一般能够获得近似于最优解。见图 6-6。

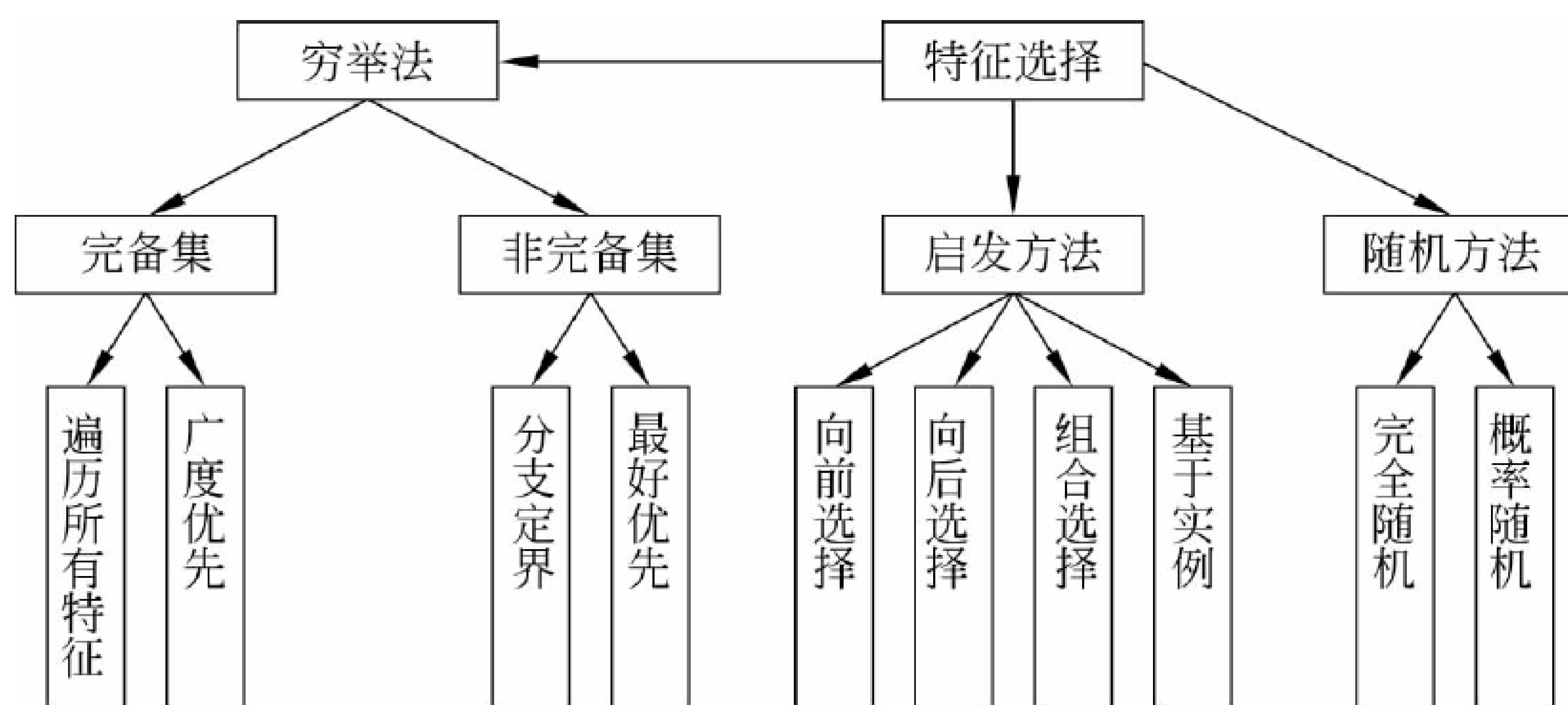


图 6-6　特征选择算法分类

随机方法是一种相对较新的方法，细分为完全随机方法和概率随机方法两种。完全随机方法是指“纯”随机产生子集，概率随机是指子集的产生依照给定的概率进行。虽然计算复杂度仍为 $O(2^N)$，但通过设置最大迭代次数可以限制复杂度小于 $O(2^N)$。常用的方法有 LVF(las vegas filter，LVF)、遗传算法、模拟退火算法及其改进方法等。这类方法需要进行参数设置，并且参数值决定是否能得到最优解。

总的来说，上述三类中穷举法能保证最优，但耗时并且计算复杂度很高，后两者以性能为代价换取简单、快速实现，但不能保证最优。实际应用中为了折中性能和代价之间的矛盾，几种方法常结合起来。

第二种是按照特征评价标准分类，特征选择可以看做是一个优化问题，其关键是建立

一种评价标准来区分哪些特征组合有助于分类,哪些特征组合存在冗余性、部分相关或者完全无关。不同的评价函数可能会给出不同的结果。根据评价函数与分类器的关系,特征选择方法分成筛选器和封装器两种。其中,筛选器的评价函数与分类器无关,而封装器采用分类器的错误概率作为评价函数。其中,筛选器的评价函数又可以细分为距离测度、信息测度、相关性测度和一致性测度。特征获取的最终目的在于使分类器的错误概率最小,因此最直观的方式是采用分类器错误概率作为评价标准,即选择使分类器的错误概率最小的特征或者特征组合。

6.6.4 文本互信息选择

互信息(mutual information,MI)在计算机模型分析中用来度量两个对象之间的相互关系,是常用的特征选择方法之一,在过滤问题中用于度量特征对于主题的区分度。

互信息本来是信息论中的一个概念,用于表示信息之间的关系,是两个随机变量统计相关性的测度,使用互信息理论进行特征抽取是基于如下假设:在某个特定类别出现频率高,但在其他类别出现频率比较低的词条与该类的互信息比较大。通常用互信息作为特征词和类别之间的测度,如果特征词属于该类,它们的互信息量最大。由于该方法不需要对特征词和类别之间关系的性质做任何假设,因此非常适合于文本分类的特征和类别的判别工作。

互信息在统计语言模型中被广泛采用,MI 越大,相似程度越大。如果 A 表示包含词条 t 且属于类别 c 的文档频数,B 包含 t 但是不属于 c 的文档频数,C 表示属于 c 但是不包含 t 的文档频数,N 表示文档总数,则可以用下面的式子来近似表示项 t 和类 c 之间的互信息:

$$\mathrm{MI} \approx \log_2 \frac{A \times N}{(A+C) \times (A+B)} \tag{6-25}$$

也可以形式化定义如下:

$$I(t,c) = \log \frac{p(t,c)}{p(t)p(c)} = \log \frac{p(t/c)}{p(t)} = \log \frac{A * N}{(A+C)(A+B)} \tag{6-26}$$

则词条 t 的平均值与最大值可以近似表示为

$$\mathrm{MI}_{\mathrm{avg}}(t) = \sum P(c)\mathrm{MI}(t,c) \tag{6-27}$$

$$\mathrm{MI}_{\max}(t) = \max_{i=1}^{n}\{\mathrm{MI}(t,c_i)\} \tag{6-28}$$

显然当 t 独立于 c 时,$\mathrm{MI}(t,c)=0$,在应用时通常取平均值或最大值。

从上式可以得出：如果 t 和 c 无关，则 $P(t/c)=0$，$I(t,c)$值就为零。如果 t 和 c 的相关性很高，并不一定 $I(t,c)$的值就很高，这与词条 t 在总文档数中出现的频数有关。在公示(6-26)中，当特征 $P(t/c)$的值相等时，由于稀有词比普通词的出现概率小，从而稀有词比普通词的分值要高，因此概率相差太大的文本特征互信息量不具有可比性，这也就使互信息在信息检索应用中具有一定的局限性。

6.6.5　χ^2 统计量特征选择

另一个常用的特征选择方法是 χ^2 统计量。在统计学中，χ^2 统计量常常用于检测两个事件的独立性。两个事件 A 和 B 独立，是指两个事件 A、B 的概率满足 $P(AB)=P(A)P(B)$ 或者 $P(A|B)=P(A)$且 $P(B|A)=P(B)$。在特征选择中，两个事件分别是指词项的出现和类别的出现。

χ^2 统计量(χ^2-statistic)的概念来自列联表检验，它可以用来衡量特征 t 和类别 c 之间的统计相关性强度，信息检索对于 χ^2 感兴趣的是那些与各个类有强关联的检索词项。则

$$\chi^2(t,c) = \frac{N \times (AD - CB)}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \tag{6-29}$$

其中 A 表示属于类别 c 并且包含特征 t 的训练文档个数；B 表示不属于类别 c 且包含特征 t 的训练文档个数；C 表示属于类别 c 且不包含特征 t 的训练文档个数；D 表示不属于类别 c 且不包含特征 t 的训练文档个数；N 为训练文档总数；n 为文档类别总数。如果 t 和 c 之间是独立的，则统计量 χ^2 的值将为 0。对于训练文本集中的每一个类，计算出每个项与该类之间的统计量 χ^2 的值。根据这些值可以求出以下两种 χ^2 的平均值或最大值：

$$\chi^2_{\text{avg}}(t) = \sum_{i=1}^{n} P(c_i)\chi^2(t,c_i) \tag{6-30}$$

$$\chi^2_{\max}(t) = \max_{i=1}^{n}\{\chi^2(t,c_i)\} \tag{6-31}$$

在一些研究中，χ^2 统计量是一种非常有效的维数约简方法。χ^2 统计方法的主要思想是：①对训练文本进行分词与索引；②在索引的基础上用统计公式计算每个词对应每个类的统计值；③选择最大值作为该词的值；④找出值最大的 N 个词作为特征项。

从统计方法的主要思想可以看出：要用统计方法计算出特征 t 和类别 c 之间的相关性，从而得出该词对类别的贡献程度大小，知道该特征 t 是否可以代表这个类别。

首先假设 t 和 c 之间符合具有一阶自由度的 χ^2 分布，再采用 χ^2 统计方法度量词

条 t 和文档类别 c 之间的相关程度。如果词条 t 对于某类的 χ^2 统计值越高，它与该类之间的相关性越大，携带的类别信息就越多，独立性也越小，则 t 对于 c 的 χ^2 值，由下式计算：

$$\chi^2(t,c)=\frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \tag{6-32}$$

因为 $N,A+C,B+D$ 均是常数，只需要关注特征词对某个类别的 χ^2 值的大小顺序，并不关心具体的值，因此把它们从式(6-32)中去掉是完全可以的，故实际计算的时候上式可以简化为：

$$\chi^2(t,c)=\frac{(AD-BC)^2}{(A+B)(C+D)} \tag{6-33}$$

特征 t 与类别 c 相互独立时，$\chi^2(t,c)=0$，此时特征值不包含任何与类别 c 有关的鉴别信息。特征 t 与类别 c 的统计相关性越强，$\chi^2(t,c)$ 的值就越大，此时特征 t 包含的与类别 c 有关的鉴别信息就越多。

χ^2 统计方法也有不足之处。首先，只考虑了特征在所有文档出现的文档频数，没有考虑特征在某一文档中出现的频率，因此对文档频率低的特征词不可靠。其次，特征词在其他类出现频率比较高，在指定类出现频率比较低时，在传统的 χ^2 统计方法中，仍然会将这些特征词作为该类的特征项。

6.6.6 基于频率的特征选择方法

基于频率的特征选择方法也就是基于文档频率的选择方法。一个特征词条 t_i 的文档频率(document frequency)是指在训练文档库中出现特征词条 t_i 的文档数。文档频率特征选择方法的基本思想是：首先设定最小和最大文档频率阈值，然后计算每个特征词条的文档频率，如果该特征词条的文档频率大于最大文本频率阈值或小于最小文档频率阈值，则删除该特征词条，否则保留。文档频率特征选择方法是基于如下假设：即如果特征词条的文档频率过小，则表示该特征词条是低频词，没有代表性。相反，如果特征词条文档频率过大，则表示该特征词条没有区分度，这样的特征词条对分类都没有多大的贡献，所以将它们删除并不会影响分类效果。

特征词条文档频率用 DF 表示，计算方法为

$$\mathrm{DF}(t_i,c_j)=\frac{\text{类别 } c_j \text{ 中包含特征词条 } t_i \text{ 的文档数}}{\text{类别 } c_j \text{ 的文档总数}} \tag{6-34}$$

基于频率的选择方法是一种简单高效的特征选择方法，相对于训练文本集规模的线性计算复杂度，能够应用于大规模训练文本库的统计。但是文档频率特征选择方法具有

如下缺点：首先文档频率特征选择方法在对特征词条进行选择操作时，认为文档频率过小的特征词条是低频词(认为它们不含有或含有很少的类别信息)，所以将它们删除并不会影响分类器的分类效果，而实际上这一假设是不全面的，存在文档频率低却能很好地反映类别信息的特征词条，文档频率特征选择方法将该类特征词条过滤掉，影响了分类器的分类效果；其次文档频率特征选择方法只考虑了特征词条是否在文档中出现，忽略了特征词条在文档中出现的频数这一重要信息。

6.7　文本的索引构建

信息检索从检索对象的内容与特征提取方面进行划分，可分为两大类，即基于文本的检索和基于内容的信息检索。文本信息检索(例如关键词检索)是目前最成熟、实践应用最成功最广泛的检索应用技术。第 5 章内容所涉及的各种信息检索一般数学原理也主要是回答基于文本的信息检索技术，而对于图形图像、视频与音频等多媒体信息的基于内容检索(例如图像的色彩、纹理、轮廓等信息内容)将在后面的章节中进行阐述和学习。

文档是按照一定结构组织的相关信息记录的集合，文档是构建各种文本型检索数据库的基础和查询的处理实体。从组织形式上划分，文档可以分为顺排文档(sequential file)和倒排文档(inverted file)两种。顺排文档就是把记录按照一定顺序完整地组织起来，在很多数据库中被称为主文档(或主文件)。例如，物品数据库依据物品记录号顺序进行建立、学生数据库依据学号顺序建立等。倒排文档就是把顺排文档中具有检索属性的项目信息抽取出来，重新排列组织成新的数据文档，在很多数据库中被称为索引文档(或辅助文件)。例如，将学生数据库中的成绩数据项抽取出来，依据学生成绩由高到低重新建立新的索引文档。索引文档是检索系统中真正具有检索意义的文档，在检索系统中由主文档生成了数据量庞大的各类索引文档。

6.7.1　基于块的排序索引方法

首先扫描一篇文档(例如一篇学位论文或一则新闻等)集合得到其中所有具有检索意义的词项，然后构造“词项－文档 ID”数据集；其次，依据词项为索引文档集的主键和文档 ID 为次键进行排序；最后将每个词项的文档 ID 组织成为倒排记录表，并计算词项频率或文档频率的统计量。对于小规模文档集来说，上述过程可以在内存中完成(例如自动词语切分与自动摘要技术)，这里的排序索引指的是大规模文档集条件下的基于块的排序索引

方法。

为了索引构建效率更高，将词项用其 ID 代替，每个词项的 ID 是唯一的序列编号。扫描原始文档集时，可以采用两遍扫描方法，第一遍扫描得到词项表，第二遍扫描构建倒排索引。

例如我们采用“桂林电子科技大学网站——新闻模块”为样本组成文档集。本文档集数据库包含 2013 年 4 月至 2016 年 4 月共三年时间跨度的桂林电子科技大学各类新闻内容，本文档集数据库大约 30GB 的数据量（新闻内容包括图像图表数据，本章内容只针对文本的索引构建进行阐述）。文档总量近 13 万篇，内容覆盖桂林电子科技大学的教学、科研、党政、招生、就业、学生工作、财务、人事、国资、后勤、基建等方面的新闻信息，实例图如图 6-7 所示。

发表时间	标 题	点 击
2016-03-31	学校召开2016年学工系统安全稳定工作布置会	1577
2016-03-31	中国运筹学会九届六次常务理事会在我校召开	843
2016-03-29	学校召开2016年国际合作与交流工作会议（图）	1812
2016-03-29	2016年我校新增3个本科专业	3020
2016-03-29	校领导赴自治区科技厅、发改委调研(图)	1714
2016-03-28	我校2012级学生陈文颖荣获国际超模大赛最佳民族服装奖	1926
2016-03-28	【综合改革】校领导带队开展学生工作模式改革专题调研（图）	1980
2016-03-27	第九届广西大学生电子设计竞赛总结暨技术交流大会在北海校区举...	1315
2016-03-27	我校4位教师荣获第十三届广西青年科技奖	1938
2016-03-27	中电二十二所吴健所长一行来校考察交流(图)	1275
2016-03-27	校党委中心组专题学习毛泽东同志《党委会工作方法》	805
2016-03-23	贵州理工学院曾羽书记一行到校考察交流（图）	1629
2016-03-23	学校召开关心下一代工作委员会2016年工作会议（图）	1067
2016-03-22	学校2016年高水平运动员招生考试工作会议召开（图）	1198
2016-03-22	【综合改革】学校召开本科专业结构调整工作布置会(图)	2169

图 6-7　桂林电子科技大学网站——新闻模块检索实例图

在统计时进行了数据的取整变换（即舍入操作），实际原始数据的总文档数 136 833 篇，每篇新闻文档的索引词项平均为 33 个，全部不同词项个数总和为 236 693 个（不包括空格和标点符号），每个词项的数据量平均为 2.3B，基于倒排索引构建的记录数为 83 996 533 个（约 8 千万个）。每个词项 ID 与文档 ID 的数据各占 2.7B，因此存储所有的词项 ID 与文档 ID 需要 2GB 的存储空间。见表 6-3。

表 6-3　桂林电子科技大学——新闻文档集统计数据表

符号	含　　义	统计值
N	文档总量	13 万篇
L	每篇文档的平均词项数量	33 个
M	词项总数	23 万
	每个词项的平均字节数(包含空格与标点)	2.7B
	每个词项的平均字节数(不包含空格与标点)	1.9B
	每个词项的平均字节数	2.3B
T	倒排记录总数	8000 万个

目前典型的数据库(例如清华同方学术期刊数据库、中国新浪网的新闻模块数据等)往往比我们这里举例(“桂林电子科技大学——新闻内容模块”为样本组成文档集,见图 6-8)的数据量要大得多,即使对于大型计算机来讲,要把全部“词项-文档 ID”集都存放在计算机内存中也是十分困难的事。如果生成的索引文件调入内存的占用空间较小,就需要进行索引压缩算法技术。对于大多数文档检索数据库而言,即使经过压缩后的倒排文档记录全部加载到计算机内存中,也是不可能的。



图 6-8　桂林电子科技大学——新闻内容文档实例

由于计算机内存空间的有限性,我们需要使用基于磁盘的外部排序算法 ESA (external sorting algorithm)。该算法的核心是:在索引排序时,尽量减少磁盘寻道次数,

因为磁盘顺序读取数据的速度要比随机寻道速度快得多。

外部排序算法 ESA 的主要基础思路是 BSBI 算法(blocked sort-based indexing algorithm),即基于块的排序算法。BSBI 步骤主要有四步。

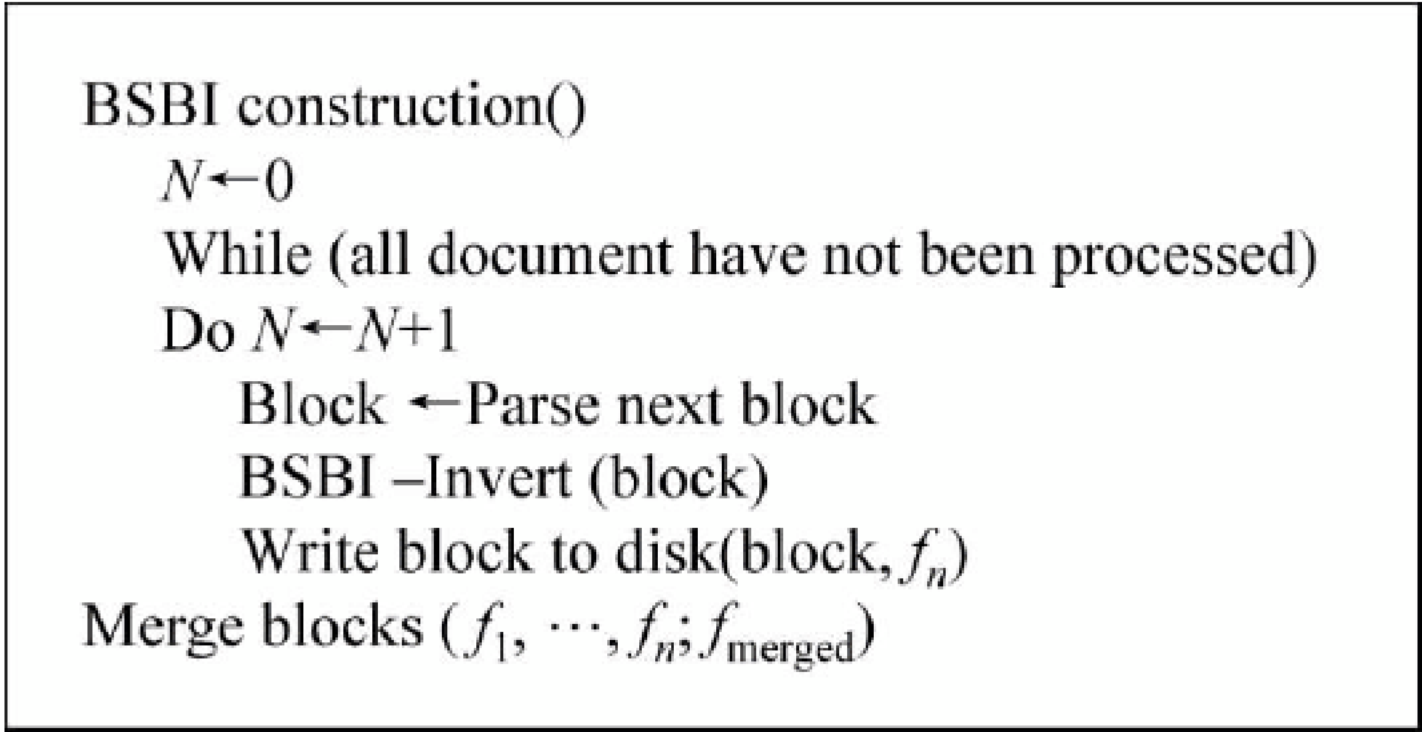
第一步,将文档集分割成多个大小相等部分。

第二步,将每个部分的"词项-文档 ID"进行排序。

第三步,将中间排序产生的临时结果存放在磁盘中。

第四步,将全部中间文档合并为最终的统一索引文档。

该算法将文档解析为"词项-文档 ID"集,并在内存中一直处理,直到累积满为一个固定大小空间为止,选择合适的块算法(见图 6-9,该算法将每个块的倒排索引存入文件 f_1, …, f_n 中,最后合并为 f_{merged}),使得文档块能够方便加载到内存中并在内存进行快速排序,排序后的块置换为倒排索引文档再写入磁盘。



```
BSBI construction()
   N←0
   While (all document have not been processed)
   Do N←N+1
      Block ←Parse next block
      BSBI –Invert (block)
      Write block to disk(block, f_n)
Merge blocks (f_1, ⋯, f_n; f_merged)
```

图 6-9 基于块的排序索引算法

建立倒排索引的过程包括:①对"词项-文档 ID"进行排序;②将具有同一词项 ID 的所有文档 ID 存放到倒排记录表中,其中每条倒排记录仅仅是一个文档的 ID;③将各个数据块索引合并为一个索引文档;④将块的倒排索引文档写入磁盘中。将该算法应用于"桂林电子科技大学——新闻"数据库,并假设内存每次能够加载 20 万个"词项 ID-文档 ID",那么算法产生 10 个索引数据块,每个数据块文档集都是倒排索引的一部分。

依据图 6-9 的算法将待合并的倒排记录表(两个数据块)从磁盘读入内存,然后在内存中合并后写入磁盘(见图 6-10)。说明:在这里为了便于理解,使用了词项本身,而不是其 ID。

BSBI 算法的复杂度主要体现在时间复杂度与空间复杂度上,在时间复杂度方面主要受排序词项数目大小、文档数据块分析时间与索引文档的合并时间的影响。

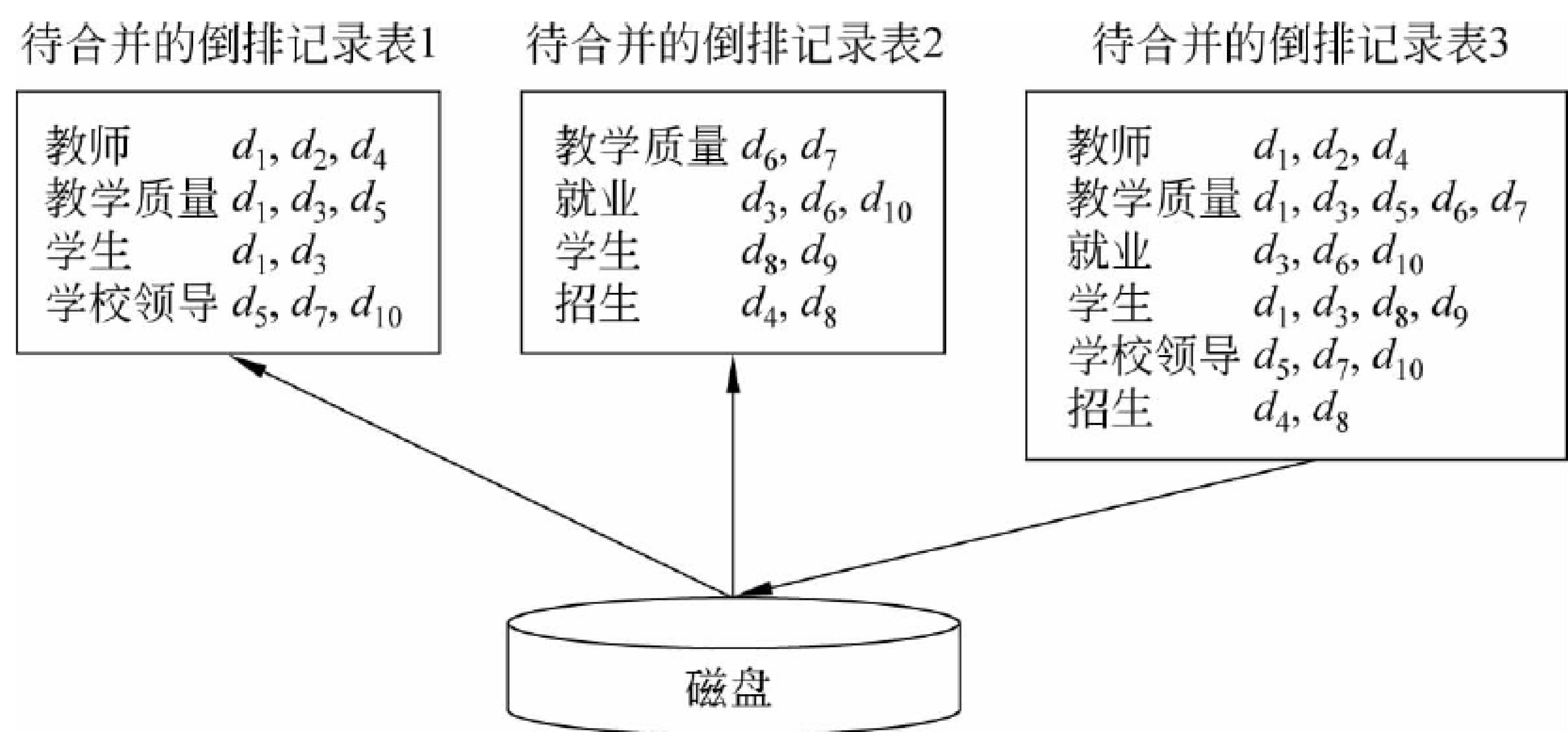


图 6-10　基于块的排序方法合并示意图

6.7.2　基于内存单次扫描的索引构建方法

基于块的排序索引方法具有很好的扩展性，但是需要一种将词项映射为与之相对应的 ID 的数据结构。对于大规模的文档集来讲，该数据结构会变得很大以至于计算机内存难以存放，一种更加有效的扩展性算法 SPIMI(single-pass in-memory indexing，基于内存单次扫描的索引算法)则能够满足这一要求。SPIMI 将每个数据块的词典(由固定文档生成的词项所组成的有序文档)写入磁盘，对于下一个块则重新采用新的词典。只要硬盘空间允许，SPIMI 算法能够构建足够大的文档索引数据库。

图 6-11 中省略了文档分析和文档转换成“词项-文档 ID”数据流，只需要循环调用 SPIMI-Invert 函数将全部文档处理完毕为止。在处理词项-文档 ID 时，如果词项是第一次出现，那么将其加入词典(由检索词项构成的数据表)，同时建立一个新的倒排记录表；如果该词项不是第一次出现，则直接返回其倒排记录表。

SPIMI 算法与 BSBI 算法的一个区别点在于，后者直接在倒排记录表中增加定位符项，且开始就需要处理形成所有的“词项-文档 ID”并进行排序；而前者是通过判定循环动态增加倒排记录表的，倒排记录表的动态处理的优势如下。

(1) 由于不需要排序(sorting)操作，数据块的处理速度大大增加。

(2) 因为保留了倒排记录表对词项的归属关系，能够大大节省内存，不需要连续保存 ID 文档。

因此，每次单独的 SPIMI-Invert 函数调用就能够处理很大的数据块，整个文档的索引构建效率也有明显提升。因为事先并不清楚每个词项的倒排记录表的大小，算法一

```
SPIMI-Invert(token-stream)
    Output-file=Newfile
    Dictionary=newhash
    While(free memory available)
    Do token ← next (token-stream)
        If term (token) ∉ dictionary
            then posting_list= addtodictionary (dictionary, term (token))
            elese posting_list= get postinglist (dictionary, term (token))
        If full(posting_list)
            Then posting_list= doublepostinglist(dictionary, term (token))
        Add to postinglist(posting_list, docID(token))
  Stored_terms ← sortterms(dictionary)
  Writeblocktodisk(sort_terms, dictionary, output_file)
Return output_file
```

图 6-11　SPIMI 算法的块倒排索引生成算法

开始会分配一个较小的倒排记录空间，如果该空间存放数据满后，就会通过判定增加新的数据块空间。当然 SPIMI 的最后一步是将多个块索引文档合并为一个完整的索引文档。

6.7.3　顺排文档索引

顺排文档索引的主要思想是将文档中的每一条记录去分别匹配用户的信息检索提问集合，文档处理完后，将提问命中结果归并后给用户。常用的顺排文档索引方法主要有表展开法、逻辑树法等。

1. 表展开法索引

表展开法是由日本学者菊池敏典 1968 年最早提出的，又称“菊池敏典算法”。该方法在信息检索的早期得到广泛应用，目前主要用于面向定题服务的检索系统。表展开法旨在将代表用户的逻辑提问式转换成检索表的形式，该检索表规定了表内容走向和检索命中与否的判断，检索时根据表内容走向及其他相关信息来判断每条记录检索是否命中。

1）表展开的含义

将经典布尔逻辑检索的逻辑提问表达式转换为逻辑检索表，每个检索词的检索组配关系要求能够用表进行精确映射，检索结果的记录是否最终命中检索需求要能准确反映出来，表展开检索能够满足这些检索要求。例如布尔逻辑表达式 $(A+B)*(C+D)$ 的展开表如表 6-4 所示。

表 6-4　$(A+B)*(C+D)$的展开检索基础表

地址	检索词	条件满足指向	条件非满足指向
1	A	3	2
2	B	3	落选
3	C	命中	4
4	D	命中	落选

表中说明了四个检索词 A、B、C、D 在地址、条件满足指向、条件非满足指向等方面的映射关系。

2）生成展开表

把逻辑检索提问式生成展开表是一个较复杂的过程，需要考虑到检索词、检索运算符、改变运算次序等内容，并生成可供检索匹配的表格形式。整个生成过程分为两部分：前处理和后处理。

（1）前处理。前处理的目的是逐个检查逻辑提问式中的字符，并从上至下填写表格。在填写表格的过程中对不同类型的表处理对象进行分别处理。

若是检索词，则将之存入展开表内的检索词栏，并记下该词在表中的地址。

若是运算符，则分别处理如下。

① 加号运算符“+”。因为两个检索词进行有“+”运算，在前一个词不满足检索条件的情况下，还可以查看后一词。当遇“+”时应在前一词的“条件不满足指向”栏中填入指向后一词的地址。

② 友好运算符“*”。如果两词进行“*”运算，在检索过程中必须均满足条件才能认为符合检索要求。当遇到“*”符号时，须在左边检索词所在行的“条件满足指向”栏中填入指向后一词的地址。

若是括号，则分别处理如下。

① 左括号“(”。将“(”后的检索词所在行的“级位”栏值加 1，同时有多级左括号时，级位值连续多次加 1。

② 右括号“)”。将“)”的紧前一个检索词所在行的“级位”栏值减 1，同时有多级右括号时，级位值连续多次减 1。

第一个检索词的级位初值为 0。在第一个检索词以后每一个检索词的初始级位由上一检索词复制得到，然后再根据条件相减。若检索词的第一个字符是左括号，则将第一个

检索词做加运算。

若遇括号结束，在最后一个检索词所在行的“条件满足指向”栏设为“命中”，“条件不满足指向”栏中设为“不命中”。

这些前处理工作至此结束，展开表中除第二、第三栏中有空白外，其余各栏均已填好，这些空白处有待于后续处理来完成填补。

（2）后处理。后处理的主要任务是按照要求设置满整个表的空白单元，设置表内容的依据是表中“级位”栏的前后级位值，填表的顺序是从下向上，直至表的顶部，从而得到一个完整的提问展开表。通常我们称表中指针所指行为“当前行”，指针移动到“当前行”之前所指向的行为上一行。

若当前行的级位值大于上一行的级位值，表示上一个的检索词后有一个右括号，如($A+B+C$)对应的检索词级位分别为1，1，0，因此，针对不同的情况应做不同处理。

若当前行的“条件不满足指向”栏为空，则表示当前行和上一行的检索词之间为“*”运算，应把上一行不满足栏内容复制到当前行的不满足栏。

若当前行的“条件满足指向”栏为空，则表示当前行和上一行的检索词之间为“+”运算，需要把上一行满足栏内容复制到当前行的满足栏。

经过上述两个处理过程，我们就可以得到一张完整的提问展开表。将若干提问式的展开表汇集起来，构成用户提问文档集合，依据用户提问文档就可以方便地进行顺排文档的检索。

3）表展开法的检索应用描述

表展开法通常用于批处理检索系统中，生成的展开表为若干逻辑提问式的集合，这个集合形成了展开表提问文档，并作为检索的提问库，专用于以后的批量检索和定题服务检索。检索时，需将所有提问展开表调入内存运行以提高查询匹配速度。检索匹配时，每从数据库中读取一条记录，就为该记录生成一个检索标识表，检索标识表由该记录的可检索项组成，然后将检索标识表中的每一检索项去匹配展开表，并对命中的检索词给以标记。当该记录标识表中的所有检索项查询完毕后，再根据每一展开表的查询情况，分析提问是否命中。对于命中者，就在相应的提问号下标注记录号及相关信息，然后再取下一条记录进行对比。全部检索匹配完毕后，才能得到本次检索的最终结果，最后通过提问号调出检索结果中各自命中结果的记录给用户。

2. 逻辑树索引

逻辑树是将逻辑提问式展开成树形结构(称主逻辑树)，运算符构成树的节点，检索词被视为树叶，所有检索词也按照有限自动机原理构造成字符树(即子树)，主树与子树间的

相关元素用指针链接。检索时,采取遍历树原则,先用文档中的标引词逐字符地遍历子树,当遍历到树的一个端头(树叶)时,依照指针标识主树,并根据遍历树方式分析提问是否命中。逻辑树展开法包括三个部分:逻辑提问式的分解、树形结构的生成、检索实现。

1)逻辑提问式分解

逻辑提问式分解的目标是提供可直接用于检索实现的主逻辑树表、检索词地址表以及检索词在检索式中的位置表。这些表在检索实践中分别发挥着各自的作用。

(1)主逻辑树表。主逻辑树表是逻辑提问式的一种树形表达形式,它用层次型的树形结构把运算符、运算项关联起来,其主要内容包括运算类型、子项个数、父项地址以及检索处理登记栏。见表 6-5。

表 6-5　主逻辑树表结构

运算类型	子项个数	父项地址	处理标识	检索处理

运算类型:用来表示逻辑提问式中的运算符类型。如“+”、“*”、“-”等,每个运算符必须有一个或多个子项,且只能有一个父项,没有父项的节点是根节点。子项个数指该运算符直接下属项的个数,下属项可以是检索词,也可以是运算符。例如 $A+B+C+D$,该运算符“+”下就有四个子项,分别为“A”、“B”、“C”、“D”。

父项地址:指本项的直接上属项在本表中的地址。如上例中的“A”、“B”、“C”、“D”都指向同一个父项“+”。

处理标识:在检索过程中填写,主要用于记录该检索项或逻辑组合项是否被“满足”。一般情况下,处理标识在检索前均为“0”,当在检索过程中被“命中”后,记为“1”,表示该项的检索过程已经完成。对于“-”运算,则处理标识栏置为 1,该词被命中后被置为“0”。

检索处理:记录该项在检索过程中的变化情况。即当该项的子项命中后,对该项进行累计处理,当该项的检索要求被满足后,就在“处理”栏置 1。例如,对于“*”运算,当其直接下属子项初次满足检索要求时,就在该栏加 1,直到该栏的数字与它的子项个数相等时,将处理标识置为 1;若为“+”运算,则当其任意一个直接下属子项初次满足检索要求时,处理标识置为 1;对于“-”运算,则在分解提问式时,就将该栏置为 1,当在以后的记录中检索到该检索词或该项的组合条件满足时,再反将其置 0,表示该项“非”运算满足。

在检索过程中,当某一行的处理标识为 1 时,就根据该行的“父项地址”值遍历到其“父项地址”行,进行检索处理,这样反复循环,当树根处(提问式的逻辑树顶端)的处理标

志为 1 时，说明该检索提问被命中。

（2）检索词地址表。检索词地址表是主逻辑树表与子表的联系纽带，在检索中当一个检索词命中以后，通过子表找到其在检索词地址表的位置，再根据该表中记录的主表位置进行检索处理。该表由两个字段组成：检索登录与检索词在主表中的位置。见表 6-6。

检索登录：该栏的作用为进行检索词命中与否的登记栏，该栏的初始值为 0，首次命中后记为 1，同时根据其在主表中的位置定位到主表，并进行检索处理。

主表位置：该词在主逻辑树表中的位置，该位置建立了主逻辑树表和子表的连接，当表中的检索词命中后，可以通过子表的指针在该表中找到主表中的相关位置。

（3）检索词位置表。检索词位置表是在逻辑提问式转换成逻辑树表的过程中，临时生成的一个中间处理过程表，该表还将作为从逻辑提问式到词逻辑树子表的桥梁，一旦子表生成完毕，该表将被清除，见表 6-7。

表 6-6　检索词地址表

检索登录	主表位置

表 6-7　检索词位置表

检索词种类	起始位置	终止位置

检索词种类：用于区别检索词的类别（如作者、关键词、标题、代码等）。设此项的目的在于区别检索对象，提高检索效率。通过种类标识分别构造检索词逻辑树表，使得在检索时可以针对不同类别的检索词去匹配不同的逻辑树。

起始位置：主要指本行检索词在整个逻辑提问式中的起始位置，以便在构造子表时，快速准确地在逻辑提问式中取词。

终止位置：指本行检索词在整个逻辑提问式中的结束位置，目的也是为了准确取词。

（4）中间工作表。从进行逻辑提问式到逻辑树表的转换过程中，由于涉及一些中间数据，这些数据在生成逻辑树时需多次使用，因此需要建立一个中间工作表来记录这些中间数据，一旦主逻辑树生成完毕，该中间工作表即可以清除，见表 6-8。

表 6-8　中间工作表结构

起始位置	终止位置	父项地址	辅助信息

起始位置：由于逻辑提问式的分解是逐层进行的，每一层可能有若干子项，使用起始位置来表示子项在逻辑提问式中的起始位置。

终止位置：记录子项在逻辑提问式中的结束位置。

父项地址：本项的父项在逻辑提问式中的地址。

辅助信息：为分解该子项时提供辅助信息。如本项的父项为何种运算，本项是否为括号项等。本算法规定："0"表示该子项的前后端分别为左右括号，"1"表示父项为"＋"，"2"表示父项为"＊"，"3"表示父项为"－"。

(5) 主逻辑树表的生成。主逻辑树表的生成算法是采用多次扫描的分层分解构造法。首先分解出逻辑式中最外层"＋"号下的子项，括号内的项暂不分解；其次扫描已分解出的子项(在最外层没有"＋"项的情况下对整个逻辑式进行)中的"＊"号的运算子项，若该子项为括号子项，则仍分解"＋"号子项；最后分解"－"号子项。

2) 逻辑树法检索应用

逻辑提问式最终转换为逻辑树的三个表：主逻辑树表、检索词地址表、检索词字符树表。这三个表构成了用户检索提问文档，整个检索也主要依赖这三个表。

实际检索过程为：从文档中读取一条记录，将记录中的标引项(主题词、责任者、分类号等可供检索的标识项)去匹配相关的检索词逻辑树，匹配成功者，根据检索词地址指针去判断检索词地址表对应的检索登录区，若为"1"，表明该词已命中过，不需再处理；若为"0"，则将该项置为"1"，同时根据本行的"主表位置"字段去修改主逻辑树表。

主逻辑树表的检索处理较为复杂，因为它不只是处理指针指向的检索词项，而且要遍历到它的父项进行相关的处理和判断。具体处理过程如下。

在主逻辑树表中该词的"处理标识"栏中填上"1"，然后根据父项地址的指针找到父项行，对"检索处理"栏做加"1"运算，再查看"处理标识"栏。若为"1"，表示该子项已做过向上遍历处理，可返回进行下一词的处理；若为"0"，则根据"运算种类"做相应处理。

若为"＋"运算，在完成标识栏置"1"，再向父项移动。

若为"＊"运算，比较"检索处理"与"子项个数"的值：如果值相等，则在完成标识栏置"1"，再向父项移动；如果值不相等，就返回进行下一词的处理。

若为"－"运算，则顺着父项进行注销处理。

随着父项指针移动到顶行时，若该行的处理标识为"1"，则表示该记录对于这一提问为命中文献信息，并将提问号和记录号写入命中文档。为了减少重复查询，实际应用时对于命中提问应采取屏蔽手段，确保该提问不再被这一记录访问处理。

与其他顺排检索方法比较，该算法虽然在分解逻辑提问式，但是在扫描次数方面可能多于其他算法例如表展开法。由于判断次数减少，其处理速度反而加快了；虽然该法对提问式的处理需要产生三个表，但它的处理是一次性的，不像展开表法分前、后处理两步；更

为重要的一点是，顺排文档检索对提问式的处理是一次性的，而且是计算机后台处理的，即使在加工提问式过程中耗费了一些时间资源，但检索处理的效率仍然较高。

6.7.4 倒排文档索引

倒排文档是相对于顺排文档而言的，是将顺排文档中可检索的信息字段项，例如信息标题、信息发布者、关键词、分类号等信息提取出来，按一定规则排序，归类相同检索项字符（例如系统的姓名字符），并把在顺排文档中相关记录的记录号集合赋予其后，以保证通过某一特征词能够快速、方便地获取相关记录信息。倒排文档方法常常又称为倒排索引。

倒排文档的组成特点，使得许多数学检索模型（如布尔模型、集合运算等）能够方便地用于文本信息检索实践中，它把两个检索词的逻辑运算转换成了两个检索词之间的记录号集合的运算。目前最常见的倒排文档检索为逆波兰展开法。

1. 倒排文档索引的建立

为了提高检索效率，希望把整个文档集合的索引都存放在内存上，但是在检索实践中这是不可行的，因为用户不希望把大部分计算机资源都用于查询工作，而且一个几百 K 数据量的文件，其全文索引的全部数据很快就达到几个 GB 的数据量。因此，对于索引工作而言，更经济的思路是把大部分索引文档存放在磁盘中，而非内存上。

倒排文档的组成元素主要包括关键字（作者、主题词、分类号等）、目长（含有该关键字记录的条数）与记录号集合（所有与该关键字有关的记录号）。倒排文档是建立在顺排文档（主文档）基础上的，它是从主文档中提取可检索字段内容，也可自动从标题、文摘或全文中自动提取关键词，利用所得到的这些属性词来建立倒排文档。

1）倒排文档的结构

倒排文档可视为主文档的辅助索引，它从不同的角度提供了对主文档的快速查询，一般来说，不同属性的数据构成不同的倒排索引文档。比如学术期刊论文在数据库中有记录号、作者、标题、关键词等属性，就可以依据作者或关键词建立索引文档。

2）倒排文档的建立

由顺排文档构造倒排文档需要经过抽词、排序、归并和组织等过程，具体实现步骤如下。

第一，选择需要构建索引的作者、关键词等字段属性，抽出其中的内容，并在其后附上其记录号。

第二，对抽出的内容进行排序，便于归并相同内容。

第三，对相同内容进行归并，把合并后的内容放入倒排文档的关键词、作者等主要字段中，统计每一数据的频次作为目长，把每一内容后的记录号顺序放在记录号集合字段。

例如，有一些学者关于信息管理方面的学术期刊论文及其基本属性见表 6-9。

表 6-9　学术论文文档及其部分属性举例

记录号	篇　　名	作者	标　引　词
1	知识管理与企业管理信息系统建设	A	知识管理，管理信息系统，企业信息化
2	论知识链与知识管理	B	知识管理，知识链，学习型组织，知识创新
3	刍议知识管理及其体系框架	C	知识管理，知识创新，知识共享
4	知识管理的组织基础	A	知识管理，学习型组织
5	论技术创新的知识空间	C	技术创新，知识空间，知识创新
6	建立企业竞争性的信息结构	A	企业信息化，信息结构，竞争情报
7	知识管理在企业竞争情报研究中的应用	B	知识管理，竞争情报，知识创新
8	管理信息系统中的文化行为研究	B	管理信息系统，企业文化
9	企业竞争情报管理系统的构建研究	C	管理信息系统，竞争情报
10	企业知识管理主体研究	C	知识管理，企业文化，管理创新

基于表 6-9 可以建立相应的关键词和作者倒排文档，见表 6-10 和表 6-11。

表 6-10　关键词索引

标 引 词	目　长	记录号集合	标 引 词	目　长	记录号集合
管理创新	1	10	学习型组织	2	2;4
管理信息系统	3	1;8;9	知识创新	4	2;3;5;7
技术创新	1	5	知识共享	1	3
竞争情报	3	6;7;9	知识管理	6	1;2;3;4;7;10
企业文化	2	8;10	知识空间	1	5
企业信息化	2	1;6	知识链	1	2
信息结构	1	6			

表 6-11　作者索引

作者	目长	记录号集合	作者	目长	记录号集合
A	3	1;4;6	C	4	3;5;9;10
B	3	2;7;8			

在建立倒排文档的过程中需要注意以下两点。

第一,倒排文档建立过程是批处理的过程,在实际的信息检索数据库建设中是不断追加检索数据的过程,因此,倒排文档的建立应具有及时更新的功能。首先,从增加的记录中取出倒排索引的字段内容;然后查询倒排索引。如果命中,则将该记录的目长加 1,并将增加记录的记录号追加进倒排文档的记录号集合字段。若没有命中,则将该字段内容以及记录号添加到倒排文档之中,并将目长置 1。

第二,由于每一个关键字所对应的记录数相差很大,因此对于只能处理定长字段的数据库或文件系统,需建立溢出文档来解决不定长问题。

2. 逻辑提问式的转换

逻辑提问式类似于算术表达式,对于信息检索而言,这种表达式并不是最优和最简洁的形式,需要进行必要的转换。1929 年波兰的逻辑学家卢卡西维兹提出了将运算符放在运算项后面的逻辑表达式,又称"逆波兰表达式"。采用这种逻辑表达式是非常方便检索运算的,日本的福岛先生最早将逆波兰表达式应用于信息检索工作,故又称为"福岛方法"。

逆波兰表达式是一种没有括号,并严格遵循"从左到右"运算的后缀式表达方法。例如,逻辑提问式"$A*(B+C)+D$"转换为逆波兰表达式就为"$ABC+*D+$",这样的表达式应用于检索将使之更加方便。因此,实现福岛方法首先要进行提问式的转换。

不论是算术表达式还是逻辑提问式中,运算符均有其运算优先级,这就决定了表达式转换具有一定的复杂度。在逻辑提问式中,其运算符的优先次序分别为:"－"、"*"、"＋",另外括号内的运算优先级最高。因此,在转换处理过程中,对运算符的优先级做如下定义(见表 6-12)。

表 6-12　运算符的优先级

运算符	优先处理的级别	运算符	优先处理的级别
(,)	1	*	3
＋	2	－	4

转换之前,需要为转换处理开辟三个存储区:用于存放转换过程中运算符的算子栈、存放检索词的检索词表存储区、存放逻辑提问式的逆波兰表达式的逆波兰输出区。

进行转换时,需从左向右逐个扫描提问逻辑式的全部字符,不同的对象做相应处理,具体如下:

(1) 遇运算符:若当前算符的优先级高于前一算符,将该算符送算子栈内;若优先级不高于(包括等于)前一算符,就将顶部算符取出送逆波兰输出区,当前算符再与栈内顶部算符比较,当前算符的优先级低就取出送逆波兰输出区,否则就将该算符送算子栈内。

(2) 遇左括号:表示其后存在一个复合检索项,暂不组成运算,应将左括号无条件置入算子栈内。

(3) 遇右括号:表示与其对应的左括号之间的所有算符都可以组成运算,栈内括号间的所有算符无条件出栈,并送逆波兰输出区,同时放弃掉这对括号。

(4) 遇运算项:将运算检索项存入检索词表,并将其在检索词表的位置送逆波兰输出区。

(5) 遇结束号:算子栈内的算子依次出栈并送入逆波兰输出区。

在转换过程中应注意两点:栈的规则是元素"后进先出",转换结束其栈为空;逆波兰输出区的算子特征为 1,检索词特征为 0。例如有一个检索逻辑表达式"$(A+B)*(C+EF)$",它的逆波兰转换处理示意如图 6-12 所示。

3. 检索指令表的生成

逻辑提问式的逆波兰表达式并不能直接用于信息检索,还需要将其转换成一组检索指令才能进行检索操作。这种转换是直接针对逆波兰表达式进行的,通过逐行扫描逆波兰输出表,根据其具体内容实现从逆波兰输出表到检索指令表的转换。操作指令表由四列元素组成:第一列为操作码,指定本行操作类型,如输入操作、运算操作、转储操作等;以后三列为操作数属性,根据操作码来决定三个操作数之间的关系,具体处理过程如下:

(1) 若为检索词,操作码置 1,第一操作数存放从逆波兰输出表中取出的检索词地址,第三操作数存放该记录号集合的工作区,例如表 6-13 将检索词表的 03 号关键词的记录号集合存放在第 2 工作区。

表 6-13　检索词操作指令表示

操作码	第一操作数	第二操作数	第三操作数
1	3		2

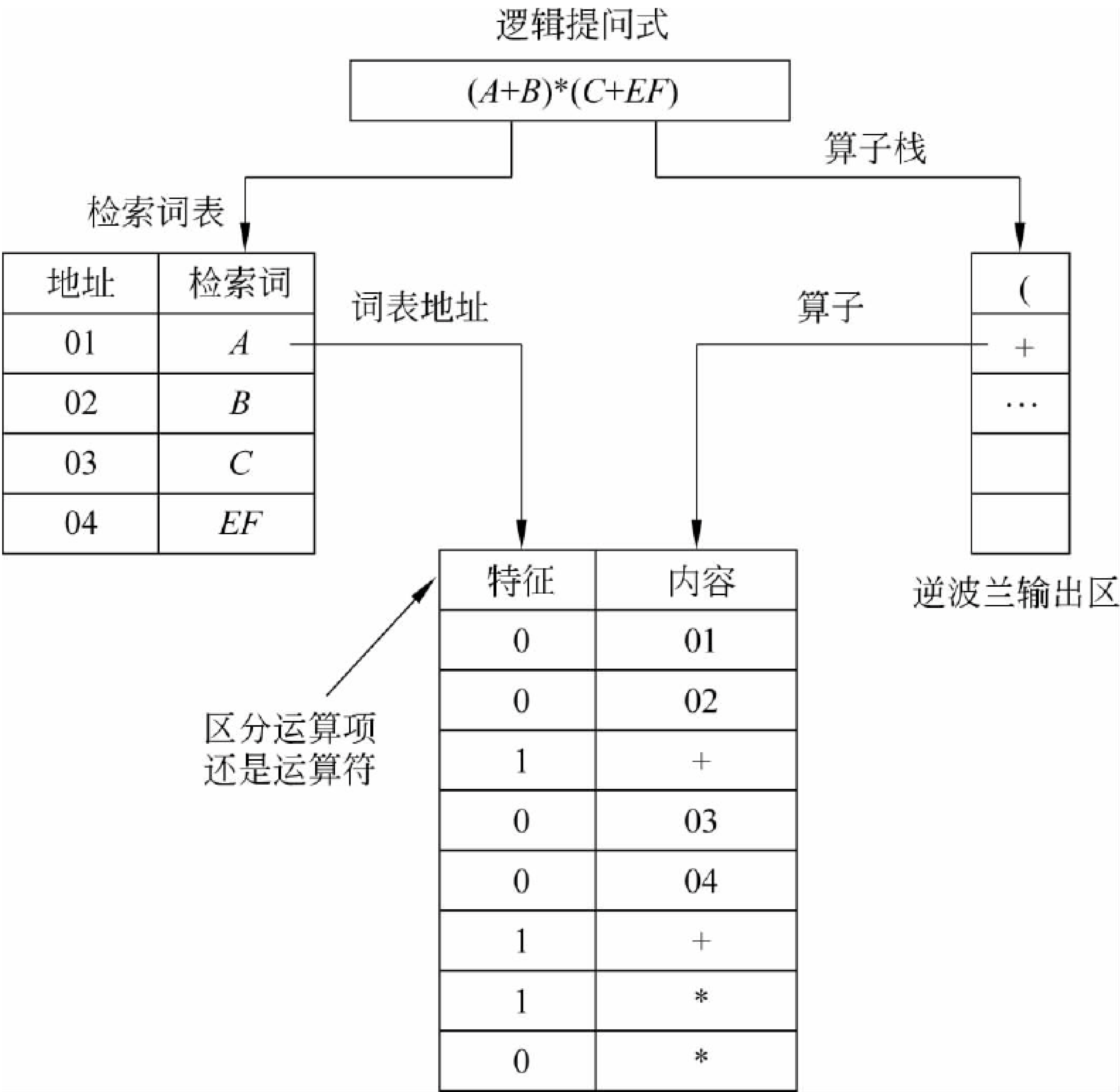


图 6-12　逆波兰转换处理实例图

若为运算符，操作码为“3”、“4”、“5”，分别代表运算符“＋”、“＊”、“－”，第一、第二操作数指定的两个工作区的记录号集合根据操作码进行相关运算，其结果送入第三操作数指定的工作区。例如表 6-14 将第 3、第 4 两个工作区的记录号集合进行“与”运算，其结果存放到第 1 工作区。

表 6-14　运算操作指令表示

操作码	第一操作数	第二操作数	第三操作数
4	3	4	1

（2）若为结束行，将操作码置 2，表示转储操作，把检索运算结果送第 7 工作区。因此，第一操作数放检索结果占用的工作区，第三操作数放置 7，表示把检索的最终结果转移到第 7 工作区。见表 6-15。

表 6-15　转储操作指令表示

操作码	第一操作数	第二操作数	第三操作数
2	4		7

(3) 转储操作结束，将最后一行的操作码置为 0，表示终止操作，其他操作数为空。见表 6-16。

表 6-16　转储操作指令表示

操作码	第一操作数	第二操作数	第三操作数
0			

由于当时计算机内存的硬件特性有限，福岛方法设定工作区为 7 个，工作区的使用从前向后遇空闲即分配，从而保证了 7 个工作区能够满足检索过程的需要。当然，7 个工作区也不是对任何形式的逻辑提问式都能满足，需要进行提问式的优化才能保证信息检索得到满足。

为了便于理解和把握关于信息检索指令表生成的基本方法，表 6-17 给出了检索式"$(A+B)*(C+EF)$"的检索指令表生成的全过程。

表 6-17　检索指令表生成过程

<table>
<tr><th>步骤</th><th>操作表状态</th><th>工作区状态表</th><th>说　　明</th></tr>
<tr><td>1</td><td><table><tr><td>1</td><td>01</td><td></td><td>1</td></tr><tr><td></td><td></td><td></td><td></td></tr></table></td><td><table><tr><td>1</td><td>1</td></tr><tr><td>0</td><td>0</td></tr></table></td><td>逆波兰表的当前检索词，操作码置 1，做"输入指令"，从上至下，第 1 工作区为空，第三操作数置 1，工作区状态表第一列表示该工作区被占用，第二列表示该工作区的运算次序</td></tr>
<tr><td>2</td><td><table><tr><td>1</td><td>01</td><td></td><td>1</td></tr><tr><td>1</td><td>02</td><td></td><td>2</td></tr><tr><td></td><td></td><td></td><td></td></tr></table></td><td><table><tr><td>1</td><td>1</td></tr><tr><td>1</td><td>2</td></tr><tr><td>0</td><td>0</td></tr></table></td><td>同上，但使用第 2 工作区，运算次序为 2</td></tr>
<tr><td>3</td><td><table><tr><td>1</td><td>01</td><td></td><td>1</td></tr><tr><td>1</td><td>02</td><td></td><td>2</td></tr><tr><td>3</td><td>1</td><td>2</td><td>3</td></tr><tr><td></td><td></td><td></td><td></td></tr></table></td><td><table><tr><td>0</td><td>1</td></tr><tr><td>0</td><td>2</td></tr><tr><td>1</td><td>1</td></tr><tr><td>0</td><td>0</td></tr></table></td><td>逆波兰表的当前行为算子，"＋"操作码为 3，表示"或运算操作"，最前空闲工作区是 3 号，所以第三操作数置 3，第 1、第 2 工作区分别放入第一、第二操作数，这两个工作区的运算结果放第 3 工作区，第 3 工作区的运算次序为 1，释放第 1、第 2 工作区</td></tr>
</table>

续表

<table>
<tr><th>步骤</th><th>操作表状态</th><th>工作区状态表</th><th>说　明</th></tr>
<tr><td>4</td><td><table><tr><td>1</td><td>01</td><td></td><td>1</td></tr><tr><td>1</td><td>02</td><td></td><td>2</td></tr><tr><td>3</td><td>1</td><td>2</td><td>3</td></tr><tr><td>1</td><td>03</td><td></td><td>1</td></tr></table></td><td><table><tr><td>1</td><td>2</td></tr><tr><td>0</td><td>0</td></tr><tr><td>1</td><td>1</td></tr><tr><td>0</td><td>0</td></tr></table></td><td>理由同第一行,运算次序为 2</td></tr>
<tr><td>5</td><td><table><tr><td>1</td><td>01</td><td></td><td>1</td></tr><tr><td>1</td><td>02</td><td></td><td>2</td></tr><tr><td>3</td><td>1</td><td>2</td><td>3</td></tr><tr><td>1</td><td>03</td><td></td><td>1</td></tr><tr><td>1</td><td>04</td><td></td><td>2</td></tr></table></td><td><table><tr><td>1</td><td>2</td></tr><tr><td>1</td><td>3</td></tr><tr><td>1</td><td>1</td></tr><tr><td>0</td><td>0</td></tr><tr><td></td><td></td></tr></table></td><td>理由同第二行,运算次序为 3</td></tr>
<tr><td>6</td><td><table><tr><td>1</td><td>01</td><td></td><td>1</td></tr><tr><td>1</td><td>02</td><td></td><td>2</td></tr><tr><td>3</td><td>1</td><td>2</td><td>3</td></tr><tr><td>1</td><td>03</td><td></td><td>1</td></tr><tr><td>1</td><td>04</td><td></td><td>1</td></tr><tr><td>3</td><td>1</td><td>2</td><td>4</td></tr></table></td><td><table><tr><td>0</td><td>0</td></tr><tr><td>0</td><td>0</td></tr><tr><td>1</td><td>1</td></tr><tr><td>1</td><td>2</td></tr><tr><td>0</td><td>0</td></tr><tr><td></td><td></td></tr></table></td><td>理由同第三行,运算次序为 2,最新占用的两个工作区 1、2 进行运算,结果放在第 4 工作区</td></tr>
<tr><td>7</td><td><table><tr><td>1</td><td>01</td><td></td><td>1</td></tr><tr><td>1</td><td>02</td><td></td><td>2</td></tr><tr><td>3</td><td>1</td><td>2</td><td>3</td></tr><tr><td>1</td><td>03</td><td></td><td>1</td></tr><tr><td>1</td><td>04</td><td></td><td>2</td></tr></table></td><td><table><tr><td>1</td><td>1</td></tr><tr><td>0</td><td>3</td></tr><tr><td>0</td><td>1</td></tr><tr><td>0</td><td>2</td></tr><tr><td>0</td><td>0</td></tr></table></td><td>理由同上,进行的是与运算操作,结果放第 1 工作区</td></tr>
</table>

续表

<table>
<tr><th>步骤</th><th>操作表状态</th><th>工作区状态表</th><th>说　　明</th></tr>
<tr><td>8</td><td><table>
<tr><td>1</td><td>01</td><td></td><td>1</td></tr>
<tr><td>1</td><td>02</td><td></td><td>2</td></tr>
<tr><td>3</td><td>1</td><td>2</td><td>3</td></tr>
<tr><td>1</td><td>03</td><td></td><td>1</td></tr>
<tr><td>1</td><td>04</td><td></td><td>2</td></tr>
<tr><td>3</td><td>1</td><td>2</td><td>4</td></tr>
<tr><td>4</td><td>3</td><td>4</td><td>1</td></tr>
<tr><td>2</td><td>1</td><td></td><td>7</td></tr>
<tr><td>0</td><td></td><td></td><td></td></tr>
</table></td><td><table>
<tr><td>0</td><td>0</td></tr>
<tr><td>0</td><td>0</td></tr>
<tr><td>0</td><td>0</td></tr>
<tr><td>0</td><td>0</td></tr>
<tr><td>0</td><td>0</td></tr>
<tr><td>0</td><td>0</td></tr>
<tr><td>1</td><td>1</td></tr>
</table></td><td>做“存储指令”和“终止指令”，第 1 工作区被释放，第 7 工作区为最终工作区，被占用</td></tr>
</table>

4. 检索实施

倒排文档可以实现对记录信息(例如图书目录数据库信息)的快速查找，查找时只需要查找索引文档就可以定位哪些信息条目与查询的请求一致。同时，利用多个关键词进行复合表达式检索时，可以在倒排文档中先完成查找的逻辑运算，获取结果后再对记录进行存取，从而提高信息检索的效率，但是要先生成检索指令表。检索指令表生成结束才真正进入实际检索处理，整个检索过程主要依赖检索词表和检索操作指令表，执行步骤按照检索指令表的顺序进行，具体操作如下。

(1) 若操作码为“1”，应进行查找和输入操作。将该行第一操作数中数据取出，根据其在检索词表中获得检索词，以该检索词去查倒排索引文档，得到的记录号集合存储到第三操作数指定的工作区中。

(2) 若操作码为“2”，说明应进行转储操作。需将第一操作数指定的工作区中的记录号集合存储到第三操作数指定的工作区中。若操作码大于“2”，表示需进行逻辑运算操作，应将第一、第二操作数指定的工作区中的记录号集合，按操作码代号进行相应的逻辑运算，运算结果存放到第三操作数工作区中。

(3) 若操作码为“0”，则表示该逻辑提问式的检索处理结束，需根据第 7 工作区的内容(命中结果)到主文档中调出命中记录，显示或打印给用户。

本章小结

信息检索从检索对象的内容与特征提取方面进行划分，可分为两大类，即基于文本的信息检索技术和基于内容的检索技术。文本信息检索技术是目前最成熟、实践应用最成功最广泛的检索应用技术。

文本分类(text categorization，TC)又称为文本自动分类，它是信息检索和文本数据挖掘的重要基础。文本自动分类能较好地解决大量文档信息归类的问题并可以应用到很多方面，如文本信息组织、文本识别、智能搜索、邮件过滤、数据挖掘、大数据处理等。文本分类的方法有决策树分类方法、*k*-最邻近分类方法、KNN 算法和朴素贝叶斯分类方法等。不同算法的精度各不相同，适用的领域也不一样。

文本分类大致经历了四个发展阶段，目前处于基于网络的大数据自动分类阶段。基于统计机器学习的文本分类技术相对成熟，被广泛应用于很多数据库检索系统或网络检索工具。其中包括基于概率方法的朴素贝叶斯分类器、基于实例的 *k* 近邻分类器、基于统计学习理论和结构风险最小原理基础上的支持向量机方法。

朴素贝叶斯分类是建立在经典的贝叶斯概率理论基础之上，其基本思想是利用特征项和类别的条件概率来估算给定文档的类别概率，是一种基于概率统计的分类方法。

多元贝努利模型(multivariate bernoulli model)或者直接称为贝努利模型(bernoulli model)。它等价于二值独立模型，对于词汇表中的每个词项都对应一个二值变量，1 和 0 分别表示词项在文档中出现和不出现。

特征选择(feature selection)是从训练集合出现的词项中选出一部分子集的过程。在文本分类过程也仅仅使用这个子集作为特征。特征选择是模式识别的关键问题之一，特征选择结果的好坏直接影响着分类器的分类精度和泛化性能。

互信息(mutual information，MI)在计算机模型分析中用来度量两个对象之间的相互性，是常用的特征选择方法之一，在过滤问题中用于度量特征对于主题的区分度。互信息在统计语言模型中被广泛采用，MI 越大，相似程度越大。

另一个常用的特征选择方法是 χ^2 统计量。在统计学中，χ^2 统计量常常用于检测两个事件的独立性。χ^2 统计方法只考虑了特征在所有文档出现的文档频数，没有考虑特征在某一文档中出现的频率，因此对文档频率低的特征词不可靠。

文档是建立各种文本型检索数据库的基础。从组织形式上划分，文档可以分为顺排文档(sequential file)和倒排文档(inverted file)两种。顺排文档就是把记录按照一定顺序

完整地组织起来，在很多数据库中被称为主文档（或主文件）。倒排文档就是把顺排文档中具有检索属性的项目信息抽取出来，重新排列组织成新的数据文档，在很多数据库中被称为索引文档。

由于索引文档数据量大，因此要考虑与索引文档运行效率紧密相关的计算机硬件参数。因为计算机内存空间的有限性，我们需要使用基于磁盘的外部排序算法，也就是基于块的排序算法思想。

基于内存单次扫描的索引算法（SPIMI）将每个数据块的词典（由固定文档生成的词项所组成的有序文档）写入磁盘，对于下一个块则重新采用新的词典。只要硬盘空间允许，SPIMI 算法就能够构建足够大的文档数据库。

顺排文档索引的主要思想是将文档中的每一条记录去分别匹配用户的检索提问集合，文档处理完毕后将各提问的命中结果归并分发给用户。常用的顺排文档索引方法主要有表展开法、逻辑树法等。

逻辑树是将逻辑提问式展开成树形结构（称主逻辑树），运算符构成树的节点，检索词被视为树叶，所有检索词也按照有限自动机原理构造成字符树（即子树），主树与子树间的相关元素用指针链接。检索时，采取遍历树原则处理。

倒排文档技术是相对于顺排文档技术而言的，是将顺排文档中可检索的信息字段项提取出来，按一定规则排序，归类相同检索项字符，并把在顺排文档中相关记录的记录号集合赋予其后，以保证通过某一特征词能够快速、方便地获取相关记录信息。倒排文档技术常常又称为倒排索引。倒排文档的组成特点，使得许多数学检索模型（如布尔模型、集合运算等）能够方便地用于文本信息检索实践中，它把两个检索词的逻辑运算转换成了两个检索词之间的记录号集合的运算。目前最常见的倒排文档检索为逆波兰展开法。

本章思考与练习题

1. 如何理解文本分类的概念含义？
2. 为什么说“文本分类技术属于一种有监督（supervised）机器学习方法”？
3. 最常见的文本分类方法是什么？
4. 国外文本分类方法经历了哪些发展阶段？
5. 完整的中文文本分类系统，一般由哪些功能模块组成？
6. 文本分类的工作过程分为哪几个阶段？ 每个阶段的任务是什么？
7. 简述朴素贝叶斯文本分类方法（NBC）的基本原理。

8．说明多项式贝叶斯分类器基本算法。

9．可以通过哪些改进技术改进朴素贝叶斯文本分类，以形成较宽松条件限制的贝叶斯网络分类器？

10．简述朴素贝叶斯分类的提升算法。

11．简述加权朴素贝叶斯文本分类原理。

12．描述基于特征相关的改进加权朴素贝叶斯文本分类的基本原理。

13．贝努利模型的基本含义是什么？

14．请比较多项式模型与贝努利模型的基本性质。

15．文本分类特征选择的含义与目的是什么？

16．特征选择方法可以分为哪些类型？

17．文本互信息的含义是什么？简述文本互信息选择的基本原理。

18．什么是 χ^2 统计量？χ^2 统计量对文本选择的作用是什么？有何不足？

19．简述基于频率的特征选择方法的基本原理。

20．数据集、训练集和测试集的各自含义是什么？

21．使用什么方法对分类器进行性能评价？

22．从哪些方面评价文本分类器的性能？

23．应用哪几个评价指标来评价文本分类器？

24．信息检索技术分为哪两大类？并各举一例说明。

25．文档的含义是什么？检索文档分为哪两大类？并各举一例说明。

26．举例说明基于块的排序索引方法原理及其主要步骤。

27．SPIMI 算法与 BSBI 算法有何区别？

28．举例说明表展开法索引的含义。

29．简述逻辑树索引的含义。

30．一般逻辑提问式最终转换为逻辑树需要哪些表？请举例说明。

31．举例说明如何建立一个倒排文档。

32．简述“逆波兰表达式”的含义。

33．利用逆波兰表达式如何构造倒排索引？请举例说明。

第 7 章　图像信息检索

近年来，随着摄像头、手机、平板、照相机等数字图像生成设备的日益普及和广泛应用，以及数字存储技术和网络通信技术的快速进步，在政治、经济、科技、军事、医学、教育、社会生活等诸多领域，每天都会产生数据量庞大的图像信息。这些数字图像中包含了大量有价值的信息，为了有效利用图像中所承载的信息价值，需要有一种能够快速而且准确地从海量图像中查找并获取所需图像的方法，也就是图像信息检索。图像检索通常分为两大类：即基于文本的图像检索和基于内容的图像检索。

基于文本的图像检索（text-based image retrieval）历史可以追溯到 20 世纪 70 年代末期。当时流行的图像检索系统是将图像作为数据库中存储的一个对象，用关键字或文本对图像进行描述。然而，完全基于文本的图像检索存在着严重的问题。首先，计算机视觉和人工智能都无法自动对图像进行标注，而必须依赖于人工对图像做出信息描述并标注。这项工作不但费时费力，而且手工的标注往往是不准确或不完整的，还常常有主观偏差。也就是说，不同的人对同一幅图像有不同的理解方法和理解角度，甚至受不同的图像理解价值取向左右，这种图像的主观理解差异将直接导致图像检索与获取的结果不准确。此外，图像中所包含的丰富视觉特征（颜色、纹理、轮廓等）往往无法用文本进行客观的描述。

20 世纪 90 年代初期，随着大规模数字图像数据库的出现，图像的准确检索与提取问题变得越来越迫切。为克服这些问题，基于内容的图像检索（content-based image retrieval）应运而生，它区别于原有系统中对图像进行人工标注的传统方法，基于内容的检索能够自动提取每幅图像的视觉内容特征作为其索引，如色彩、纹理、形状等图像内容特征，这种方法从一个新的视角建立了图像检索的整体框架。

7.1　图像基础知识

为了更好地理解图像检索基础知识与基本原理，首先需要掌握有关图像的一些基本知识，包括图像色彩的要素、图像属性类型与图像格式方面的知识。

7.1.1 图像色彩三要素

图像色彩三要素指的是色彩亮度、色调与饱和度，是色彩在视觉上的反映特性，人眼得到的任何颜色都是这三种要素的综合效果与整体结果。其中色调与光波的波长有直接关系，亮度和饱和度与光波的幅度有关。

（1）亮度。亮度是光作用于人眼时所引起的明亮程度的感觉，它与被观察物体的发光强度有关，在色彩上反映为色彩的明暗程度，例如深红色和浅灰色就与亮度相关。亮度有时称为明度，计算明度的基准是灰度测试卡。黑色为 0，白色为 10，在 0～10 之间等间隔地排列为九个阶段，如图 7-1 所示。色彩可以分为彩色和非彩色，但后者仍然存在着明度。作为彩色，每种色各自的亮度、暗度在灰度测试卡上都具有相应的位置值。

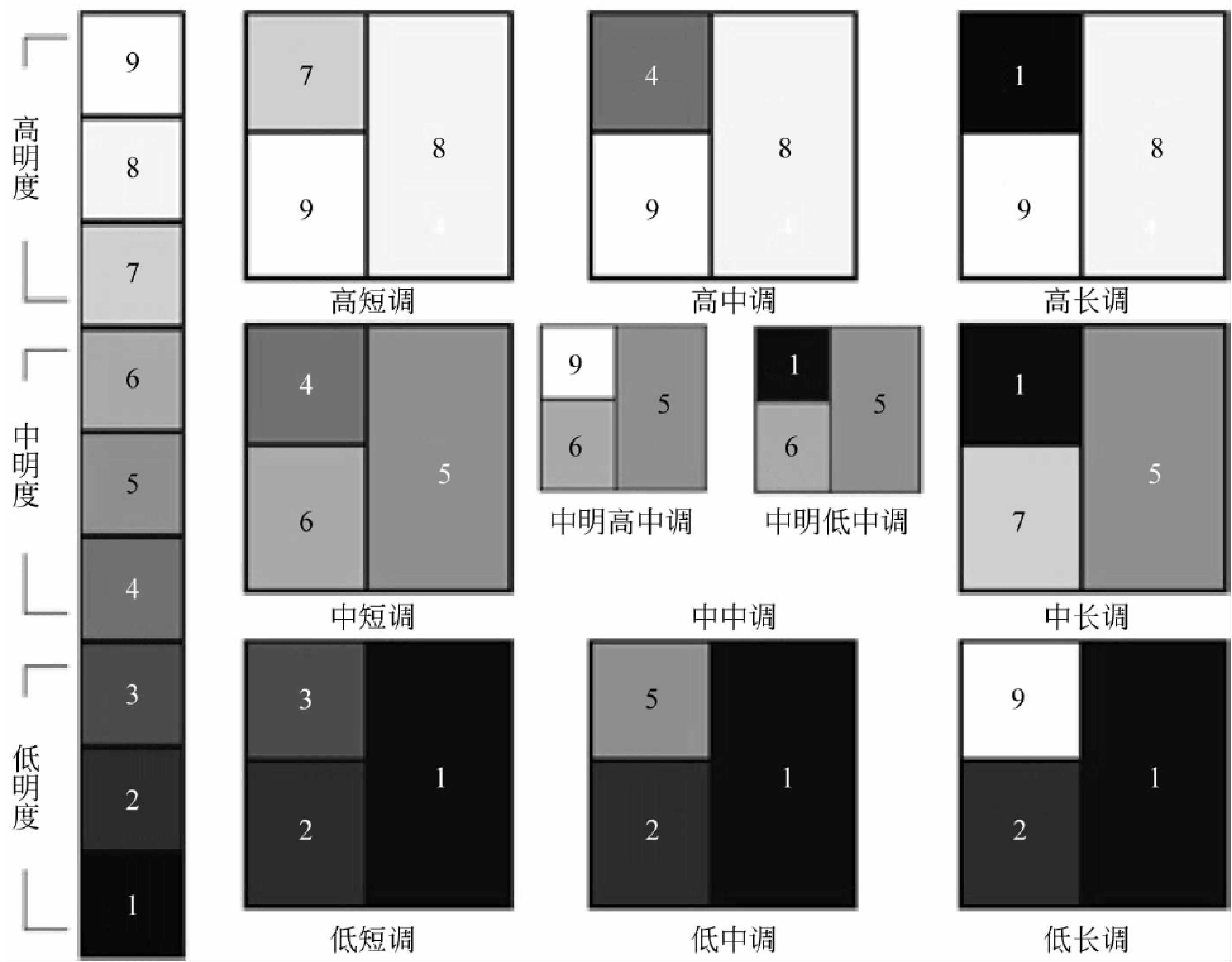


图 7-1 色彩亮度的灰度测试卡示意图

（2）色调。色调是当人眼看到一种或多种波长的光时所产生的色彩感觉，它反映颜色的种类，决定颜色的基本物理特征，如红色和棕色就是指色调。色调有时称色相，颜色

的不同是由光的波长的长短差别所决定的。作为色相,指的是这些颜色不同波长的情况。波长最长的是红色,最短的是紫色。把红、橙、黄、绿、蓝、紫和处在它们各自之间的红橙、黄橙、黄绿、蓝绿、蓝紫、红紫这六种中间色,共计 12 种色作为色相环(如图 7-2 所示)。

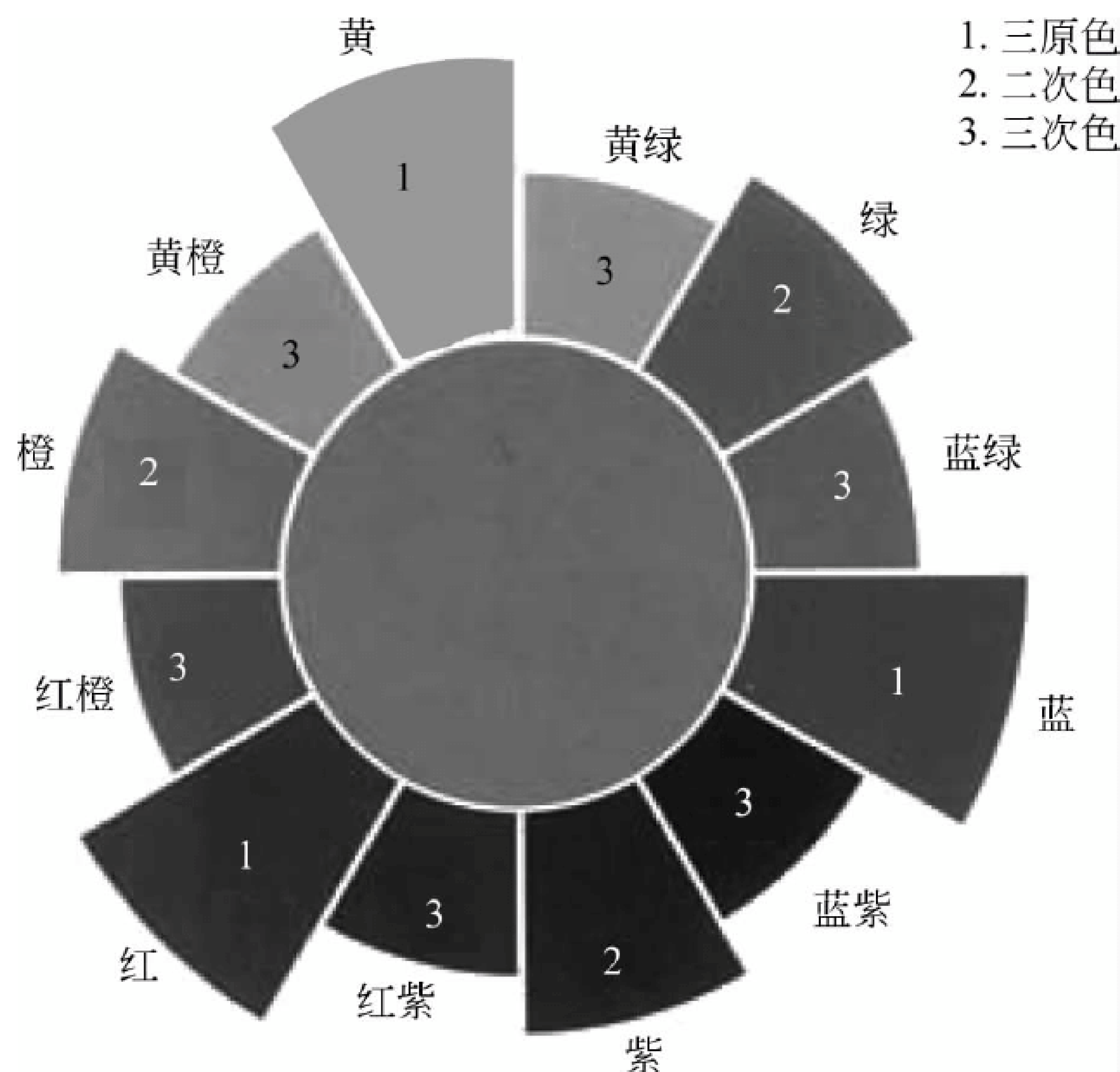


图 7-2　色调环形构成示意图

在色相环上排列的颜色是纯度高的颜色,被称为纯色。这些颜色在环上的位置是根据视觉和感觉的相等间隔来进行组织的。用类似这样的方法还可以再分出差别细微的多种颜色来。在色相环上,与环中心对称,并在 180 度位置两端的色被称为互补色。人眼得到的可见光是由红、橙、黄、绿、青、蓝、紫七种颜色组成的,波长在[380nm,740nm]之间。其中红色、绿色和蓝色为“三原色”(或称做三基色,RGB),三原色的含义是这三种颜色依据不同比例可以组成其他任何类型色彩。

(3) 饱和度。饱和度指的是颜色的纯度,即掺入白光的程度,或者说是指颜色的深浅程度。对于用于色调的彩色光,饱和度越高,颜色越鲜明(或者说越纯)。通常把色调和饱和度通称为色度,用数值表示颜色的鲜艳或鲜明的程度称之为彩度。彩色的各种色都具有彩度值,无彩色的彩度值为 0,对于彩色的彩度(纯度)的高低,区别方法是根据这种彩色中含灰色的程度来计算的。彩度由于色相的不同而不同,而且即使是相同的色相,因为明度的不同,彩度也会随之变化。

7.1.2 图像的三种基本类型

（1）位图图像。位图图像（bitmap），也称为点阵图像，是由称做像素（图片元素）的单个点组成的。这些点可以进行不同的排列和染色以构成图样。当放大位图时，可以看见赖以构成整个图像的无数单个方块。扩大位图尺寸的效果是增大单个像素，从而使线条和形状显得参差不齐。位图文件记录了图形或图像的每一个像素点的位置及代表该像素颜色的数值等信息。一般来讲，同一位图构成的像素点越多，图像越清晰，例如同样一张风景图像，800 万像素就比其 300 万像素清晰得多。根据有损压缩和无损压缩方法的最后结果，该类型图像又有多种格式，例如，.bmp 图、.tif 图、.gif 图、.jpg 图等。

（2）矢量图图像。矢量图，也称为面向对象的图像或绘图图像，在数学上定义为一系列由线连接的点。矢量图像文件中的图形元素称为对象。每个对象都是一个自成一体的实体，它具有颜色、形状、轮廓、大小和屏幕位置等属性，是计算机通过数学运算而产生的图形，而不是像位图那样逐点描述的，因此，该图形所占容量很小，而且它的显示效果不受图形大小或显示器分辨率的影响。矢量图的文件格式因生成它的软件的不同而不同。矢量图形的格式也很多，例如，Adobe Ⅲ ustrator 生成的 *.AI、*.EPS 和 SVG 图，AutoCAD 的 *.dwg 和 dxf，Corel DRAW 的 *.cdr，Windows 标准图元文件 *.wmf 和增强型图元文件 *.emf 等。

（3）印刷图。印刷用图片不同于平常计算机显示用 RGB 图片，必须为 CMYK 模式。CMYK 代表印刷上用的四种颜色：C 代表青色，M 代表洋红色（也称品红），Y 代表黄色，K 代表黑色。印刷用图片输出时将图片转换为网格点，也就是 dpi（dots per inch，每英寸的点数量）精度。印刷用图片在精度上理论最小值要达到 300dpi。传统胶印采用的都是柯氏印刷（四色套印），也就是将彩色图片分成青（C）、品（M）、黄（Y）、黑（B）四色网点，再晒成 PS 版，经过胶印打印机四次印刷，出来后就是彩色的印刷成品。

7.1.3 常用图像文件格式

在实际的图像检索活动中，大量的图像文件在格式上是多种多样的。通常在图像数据库中，存取的图像格式也不统一，因为图像生成或者产生的途径与形式本来就是多样的。

（1）BMP（bitmap picture，位图）。BMP 图像文件格式是一种 Windows 或 OS2 标准的位图式图像文件格式，它支持 RGB、索引颜色、灰度和位图样式模式，但不支持 Alpha 通道。该文件格式还可以支持 1～24 位的格式，其中对于 4～8 位的图像，使用 Run

Length Encoding(RLE 为运行长度编码)压缩方案,这种压缩方案不会损失数据,该格式非常稳定,在文件大小没有限制的场合中运用极为广泛。这种格式的特点是包含的图像信息非常丰富,几乎不进行压缩,也由此导致了它与生俱来的缺点就是占用空间较大,因此,目前 BMP 在单机上比较流行。

(2) GIF(graphics interchange format,图形交换格式)。GIF 图像是一种无损耗的图像格式,在各种平台的图形处理软件上均可处理的经过压缩的图形格式。GIF 是一种布尔透明类型,它既可以是全透明,也可以是全不透明,但是没有半透明的属性。GIF 使用了一种叫做 LZW(lempel-ziv-welch encoding,LZW),即字符串表压缩算法进行压缩,在 GIF 的压缩过程中,像素是由上到下水平压缩的,这也意味着同等条件下,横向的 GIF 图片比竖向的 GIF 图片更小,它不适合照片,但适合对颜色要求不高的图形(比如说图标、图表等)。GIF 支持动画,目前网络上大量采用的彩色动画文件多为这种格式的文件,也称为 GIF 动画格式文件。此外考虑网络传输中的实际情况,GIF 图像格式还增加了渐显方式,即在图像传输过程中,用户可以先看到图像的大致轮廓,然后随着传输过程的继续而逐步看清图像中的细节部分,从而适应了用户的"从朦胧到清楚"的观赏心理。GIF 不能存储超过 256 色的图像。

(3) JPEG(joint graphic expert group,联合图像专家组)。JPEG 是可以大幅度地压缩图形文件的一种图形格式,JPEG 是一种有损压缩格式,此格式的图像通常用于图像预览和一些超文本文档中(HTML 文档)的图像嵌入。因此在普通应用领域(非特殊要求领域),该格式的图像普及率与流行度最高,例如,各种手机、数码照相机所采集的大多数图像均为 JPEG 格式。JPEG 格式的最大特色就是文件比较小,可以进行高倍率压缩,是目前所有格式中压缩率最高的格式之一。JPEG 格式存储的文件数据量是其他类型的图形文件的 1/10～1/20,而且色彩数最高可达到 24 位,因此被广泛应用于网络上的网页或 Internet 上的图片库。JPEG 格式在压缩保存的过程中会以矢量最小的方式丢掉一些肉眼不易察觉的数据,因此保存的图像与原图有所差别,没有原图的质量好,因此印刷品最好不要用此图像格式。

(4) TIFF(＊.tif)。TIFF 的英文全名是 tagged image file format(标记图像文件格式),是 Mac 中广泛使用的图像格式,它由 Aldus 和微软联合开发,最初是为跨平台存储扫描图像的需要而设计的。该格式分为有损压缩和无损压缩两种形式,最高支持的色彩数可达 16 位。TIFF 格式存储信息量大,便于应用程序之间和计算机平台之间图像数据交换,细微层次的信息较多,有利于复制,但文件体积大,图像格式复杂。该格式的压缩方式可采用 LZW 无损压缩方案存储。

(5) PSD(＊. PSD)。PSD 格式是 Adobe Photoshop 软件自身的格式,这种格式可以存储 Photoshop 中所有的图层、通道、参考线、注解和颜色模式等信息。PSD 其实是 Photoshop 进行平面设计的一张草稿图文件或者工程图文件,里面包含有各种图层、通道、遮罩等多种设计的样稿,以便于下次打开文件时可以修改上一次的设计。PSD 格式所包含图像数据信息较多(如图层、通道、剪辑路径、参考线等),大多数排版软件不支持 PSD 格式文件。

(6) PNG(portable network graphics,便捷网络图)。PNG 是一种新兴的网络图像格式。1994 年年底,由于 Unysis 公司宣布 GIF 拥有专利的压缩方法,要求开发 GIF 软件的作者必须缴纳一定费用,由此促使免费的 PNG 图像格式诞生。1996 年 10 月 1 日由 PNG 向国际网络联盟提出并得到推荐认可标准,大部分绘图软件和浏览器开始支持 PNG 图像浏览。PNG 是目前保证图像信息最不失真的格式,PNG 格式包括许多子类,存储形式丰富,兼有 GIF 和 JPEG 的色彩模式,在实践中大致可以分为 256 色的 PNG 和全色的 PNG,用户可以用 256 色的 PNG 代替 GIF,用全色的 PNG 代替 JPEG。PNG 的一个特点是能把图像文件压缩到极限以利于网络传输,又能保留所有与图像品质有关的信息。PNG 的另一个特点是支持间隔渐进显示,显示速度很快,只需下载 1/64 的图像信息就可以显示出低分辨率的预览图像,但是会造成图片变得更大。PNG 同样支持透明图像的制作。PNG 的缺点是不支持动画应用效果,PNG 有 GIF 的所有特点,但比 GIF 更具有优势的是它支持 alpha 透明和更优的压缩。

(7) SVG(scalable vector graphics,可缩放的矢量图)。它是由 World Wide Web Consortium(W3C)联盟进行开发的基于 XML 应用的图像格式。严格来说,应该是一种开放标准的矢量图形语言,可让用户设计高分辨率的 Web 图形页面。用户可以直接用代码来描绘图像,可以用任何文字处理工具打开 SVG 图像,通过改变部分代码来使图像具有交互功能,并可以随时插入到 HTML 中通过浏览器来观看。SVG 提供了目前网络流行格式 GIF 和 JPEG 无法具备的优势:可以任意放大图形显示,但绝不会以牺牲图像质量为代价;只在 SVG 图像中保留可编辑和可搜寻的状态;SVG 文件比 JPEG 和 GIF 格式的文件要小很多,因此下载也很快。SVG 的开发将会为 Web 提供新的图像标准。

(8) EPS(encapsulated PostScript)。EPS 是 PC 用户比较少见的一种格式,而苹果 Mac 的用户则用得较多。它是用 PostScript 语言描述的一种 ASCII 码文件格式,主要用于排版、打印等输出工作。用 PostScript 语言描述的 ASCII 图形文件,在 PostScript 图形打印机上能打印出高品质的图形图像,其最大的优点是可以在排版软件中以低分辨率预览,而在打印时以高分辨率输出。

(9) CDR(CorelDraw)。CDR 是 CorelDraw 软件工具的文件格式。CDX 是所有 CorelDraw 应用程序均能使用的图形图像文件,是发展成熟的 CDR 文件。

(10) DIB(device independent bitmap)。描述图像的能力基本与 BMP 相同,并且能运行于多种硬件平台,只是文件较大。

(11) DIF(drawing interchange format)。AutoCAD 中的图形文件,它以 ASCII 码方式存储图形,表现图形的大小方面十分精确,可以被 CorelDraw,3d MAX 等大型软件调用编辑。

(12) DXF(drawing exchange format)。DXF 是 AutoCAD 中的矢量文件格式,它以 ASCII 码方式存储文件,在表现图形的大小方面十分精确。许多软件都支持 DXF 格式的输入与输出。

(13) EMF(enhanced metafile)。EMF 是微软公司为了弥补使用 WMF 的不足而开发的一种 Windows 32 位扩展图元文件格式,也属于矢量文件格式,其目的是使图元文件更加容易接受。

(14) IFF(image file format)。用于大型超级图形处理平台,比如 AMIGA 机,好莱坞的特技大片多采用该图形格式处理。图形(图像)效果,包括色彩纹理等逼真再现原景。当然,该格式耗用内存、外存等计算机资源较大。

(15) FLIC(FLI/FLC)。FLIC 格式由 Autodesk 公司研制而成,FLIC 是 FLC 和 FLI 的统称。FLI 是最初的基于 320×320 分辨率的动画文件格式,而 FLC 则采用了更高效的数据压缩技术,具有比 FLI 更高的压缩比,其分辨率也有了不少提高。

(16) MPT(macintosh paintbrush)或 MAC。Macintosh 机所使用的灰度图像模式,在 Macintosh Paintbrush 中使用,其分辨率只能是 720×567。

(17) PCD(Photo CD)。由柯达公司开发,其他软件系统对其只能读取。

(18) PCP(PC paintbrush)。由 ZSoft 公司创建的一种经压缩且节约磁盘空间的 PC 位图格式,最高可表现 24 位图形图像。过去有一定市场,但随着 JPEG 的兴起,其地位已是日薄西山。

(19) PCX。PCX 格式是由 ZSoft 公司在开发图像处理软件 paintbrush 时开发的一种格式,这是一种经过压缩的格式,占用磁盘空间较少。由于该格式出现的时间较长,并且具有压缩及全彩色的能力,所以现在仍比较流行。

(20) TGA(tagged graphics)。TGA 文件格式是由美国 Truevision 公司为其显卡开发的一种图像文件格式,已被国际上的图形图像工业所接受。TGA 的结构比较简单,属于一种图形图像数据的通用格式,最高色彩数可达 32 位。VDA、PIX、WIN、BPX、ICB 等

均属旁系。TAG格式在多媒体领域有着很大的影响,是计算机生成图像并向电视转换的一种首选格式。

(21) WMF(windows metafile format)。WMF是windows中常见的一种图元文件格式,属于矢量文件格式。它具有文件短小、图案造型化的特点,整个图形常由各个独立的组成部分拼接而成。该类图形比较粗糙,并且只能在Microsoft Office中调用编辑。

除此之外,Macintosh机专用的图形图像格式还有PNT、PICT、PICT2等。

7.2 图像检索概述

图像的数据库传统管理方式是以文件系统为中心进行展开的,当用户查询一幅图像时,要逐一打开文件进行浏览才能找到其目标图像,随图像文件数量的急剧增加,查找效率直线降低。由于以文件存储方式对图像的使用和操作非常方便,因而以文件管理图像的方式一直延续至今。基于图像内容(形状、纹理、颜色等)的检索技术则能够克服基于文本形式的图像检索的一些重要缺陷以提高其检索精度。

7.2.1 图像检索一般模型

图像检索一般模型(见图7-3)主要包括以下几个方面的内容。

(1) 图像特征提取。图像特征提取一般从两个方面入手,图像底层特征提取和语义特征提取。底层特征一般包括图像的色彩、纹理和形状特征,其特点是这些特征对于指定图像是唯一的、定量的。图像语义特征比较难提取,目前通过人工提取或者人机交互来获得图像语义特征。

(2) 检索匹配机制。针对基于色彩特征的图像检索,常用的检索匹配机制有直方图距离、欧氏距离、信息熵等,至于哪种匹配机制最有效,并没有严格意义的定论。在检索系统中,合理选取检索匹配机制是十分重要的,很多时候需要多种检索匹配机制联合工作才能取得较好的效果。

(3) 检索者终端。检索者终端是指用户与系统的接口,包括索引机制和反馈机制。索引机制包括按例查询(query by example,QBE)和按草图查询(query by sketching,QBS)。一般来说,QBE是现在基于图像内容检索的必备索引机制,而QBS对检索者的要求较高,大多数检索者不会手工绘图进行查询。

(4) 相关反馈(relevant feedback)。相关反馈是指检索者对于检索结果的反馈,检索系统会根据相似度排序给出检索结果,检索者则可以通过反馈系统对检索结果进行信息

反馈,系统会动态调整相关权重或其他参数,以此来完成二次检索,甚至多次检索来达到图像检索需求的满足。

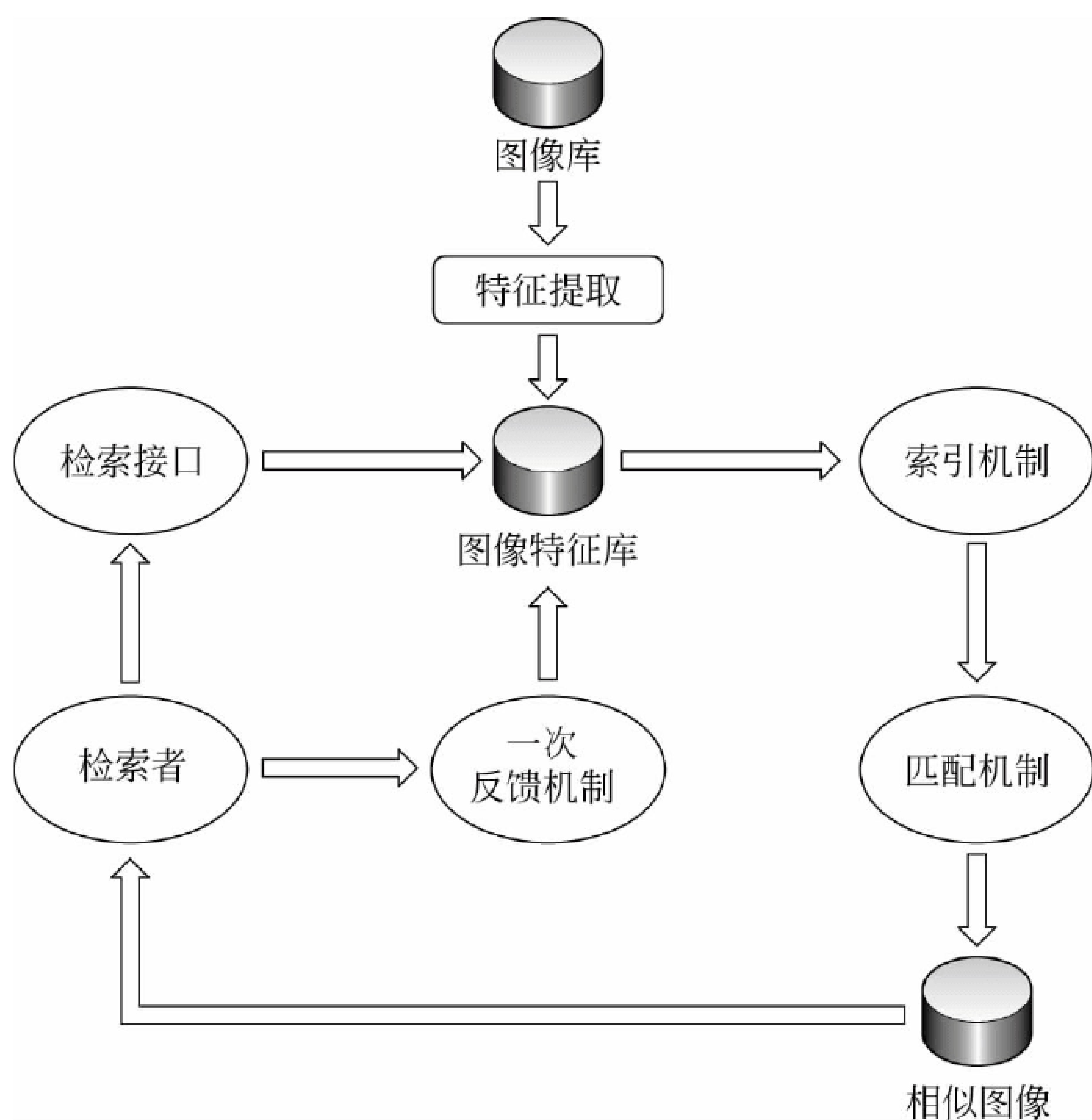


图 7-3　图像检索一般模型

7.2.2　基于文本方式的图像检索

早在 20 世纪 70 年代,数据库专家就开始研究如何对图像数据进行有效的管理,其主要方法是对图像文件建立关键词或图像标题以及一些附加描述信息,然后在图像的存储路径和图像关键词之间建立联系,传统的关系数据库技术就可以满足这样的要求。现在数据库技术已经取得了长足的进步,一些大的商用数据库系统都开始支持以二进制数据形式存储图像,但对图像的管理仍是通过二进制数据形式和图像的关键词建立联系来实现的。由于直接在数据库中访问图像的操作比较复杂,因此在数据库中以二进制数据形式管理和检索图像的方法在目前并没有流行起来。

7.2.3 基于知识和视觉特征的图像检索

事实上,对图像视觉特征进行管理在20世纪70年代就曾经引起了人工智能和模式识别等领域的关注,并取得了一定的成就。这时的图像数据库主要是应用在某一特定的领域,往往和其他信息系统结合在一起使用,主要涉及地理信息系统、病人X照片的归档、检索和诊断系统,以及人脸识别和指纹识别系统。在70年代到80年代初,采用关系数据库子系统和图像存储管理子系统集成设计成图像数据库系统,对图像数据进行检索,图像检索主要包括属性检索、结构检索、相似检索以及这几种方式的综合检索。REDI是普度大学完成的一个综合数据库系统,它与一个图像数据理解系统之间保留有接口,该系统通过图像处理和模式识别方法提取出图像的结构信息和特征,查询操作采用关系查询语言,它涉及空间关系和常规的查询。在随后又出现了用二维符号串(2D-string)来表达一幅逻辑图像的空间关系,并将此方法用于图像检索系统中。

早期图像数据库的典型应用是地理信息系统,随后一些人工智能研究者在研究和开发专家系统的过程中,采用图像数据来加强对问题的解释能力,运用了图像的一些模式特征,并对这些特征进行一定的语义解释,例如采用图像数据库技术来管理病人的心脏照片。在现在的指纹识别系统和人脸的照片管理系统中已经取得了较成功的运用。早期的图像数据库规模小且仅应用在特定的领域,检索方面也大都以精确模式匹配为主。

7.2.4 基于内容的图像检索

20世纪80年代是多媒体技术发展的时代,图像的获取、创作、压缩、存储技术都取得了举世瞩目的成就,而对图像信息的检索应用尚未给予足够的重视。90年代是计算机网络时代,特别是90年代中期以来以Web为代表的信息发布以及资源访问方式的广泛流行,信息的发布方式也从单一文本方式转变为以图形、图像、动画、视频和音频等视听信息为一体的多媒体方式。整个Internet网络环境就像一个大型的分布式数据库,在其中寻找自己感兴趣的任何一种媒体信息犹如大海捞针,因此对网络信息检索工具的依赖日益加强。而目前基于网络的检索工具如Baidu、Google、Yahoo、Info seek和Lycos等大多采用基于文本检索的方式去获取图像文件,这种采用对图像建立关键词等文本描述图像信息的方式已越来越不适应网络信息检索的要求。基于文本的图像检索主要存在以下局限。

(1) 对图像标识文本信息仍由手工完成,随着图像数据来源日益广泛,这种方法显得费时费力。

(2) 文本描述信息是非常主观的,不同的人对同一幅图像数据可能有不同的理解,因此当用户在查询时输入的关键词和数据库中的关键词不一致或这些关键词根本就不存在时,将导致检索失败。

(3) 采用关键词形式很难将图像所反映的内容描述清楚并描述完整,因为“一幅画胜过千句话”。

(4) 由于媒体信息是发布在 Internet 网络环境中,不同国家不同民族很难用同一种语言对图像进行标识和描述,而且对图像语义理解的差异性也很大。

为了突破文本检索方式的诸多弊端,人们又转向研究图像中所包含的内容信息作为图像的索引,对这方面的研究要归功于模式识别研究者,其主要的方法是根据图像的色彩、纹理、图像对象的形状以及它们的空间关系等内容特征作为图像的索引,计算查询图像和目标图像的相似距离,按相似度匹配进行检索,其目的是试图解决图像数据库系统中手工建立文本标识信息的诸多缺陷。

作为传统数据库检索的拓展,基于内容的图像检索系统主要是根据图像的内容进行检索。同传统的关系数据库检索系统相比,它主要具有以下特点。

(1) 传统的数据库中,符号数据可以用基本数据类型精确地表示,检索匹配是精确匹配。而图像数据是一段二进制数据流,对图像进行像素和像素的精确匹配不科学。事实上人对两个图像的相似和不相似的判断是根据图像中所包含的内容,很难将其精确描述,因此内容的表达是近似的。

(2) 图像数据的表达不是单一的,多种表达方法并存是必要的,表达方法的选择要依赖于特定的用户和特定的应用领域,随着识别技术的发展还可能采用更新或更好的表达方法。

(3) 符号数据本身就具有语义信息,在符号数据命名的过程中就赋予了特定的信息。图像中的内容本身不包含语义信息,对图像的匹配主要是对图像中的内容特征进行相似匹配。

(4) 由于对内容表达不精确,因此检索得到的结果可能包含一些不相关的图像,这种情况对基于内容的检索是允许的,但重要的一点是在检索中不要将相关的图像过滤掉。

7.2.5　图像内容描述的标准化

由于基于内容的图像检索有着广泛的需求和较好的市场前景,因而也引起了国际标准化组织的关注,MPEG(动态图像专家组)正在着手制定更高版本的 MPEG-7(又称为多媒体内容描述接口),它主要是对各种类型的多媒体数据进行规范化描述,目的是便于快

速和有效地查找用户感兴趣的图像资料。MPEG-7 的推出将产生广泛的应用前景，包括数字图书馆、多媒体目录服务、广播媒体的选择、多媒体编辑等。这些潜在的应用将对下面的应用领域产生巨大的影响，如教育、娱乐、调查服务、地理信息系统、医疗应用、电子购物、电影、视频和无线广播归档等。随着多媒体内容描述的标准化，图像内容的描述也将随之而标准化，基于内容的图像检索将朝商业化方向快速迈进。

综上所述，对图像的存储与检索早期是采用文件方式；在 20 世纪 70 年代到 80 年代期间是采用关键词等描述方法建立图像的索引，这个时期主要以数据库学派的研究为主，同时出现了以视觉特征为图像索引的面向特定应用的小规模图像数据库系统；90 年代以后，人们转向研究以面向网络环境支持基于内容检索的大规模图像数据库系统，这个时期主要以模式识别学派的研究为主。到 2000 年以后随着 MPEG-7 的推出，图像检索将朝标准化和商业化方向快速发展。

7.3 基于图像内容特征提取

图像特征提取是基于内容的图像检索的基础，广义上讲，特征应该包括图像的文本特征（图像名称、关键词、注释等）和图像视觉特征（颜色、纹理、形状等）。视觉特征可以进一步分为通用特征和领域相关特征，前者包括颜色、纹理以及形状特征；后者与具体的应用紧密相关，如人的面部特征和指纹特征等。由于感知的主观特性，对于给定的特征并不存在一种最佳的表达方式，图像特征的不同表达方式从各个不同的角度刻画了该特征的某些性质。

7.3.1 基于颜色特征的图像检索

在图像检索中颜色特征是应用最广泛的视觉特征，它在复杂背景和不依赖于图像的大小和方向时应用较多。图像颜色特征是一种全局特征，描述了图像或图像区域所对应的景象的表面性质。一般颜色特征是基于像素点的特征，此时所有属于图像或图像区域的像素都有各自的贡献。由于颜色对图像或图像区域的方向、大小等变化不敏感，所以颜色特征不能很好地捕捉图像中对象的局部特征。在颜色特征方面，颜色直方图描述了图像颜色的统计分布特征且具有平移、尺度、旋转不变性，因此通常用颜色直方图来描述颜色特征。

1. 颜色模型

计算机系统中，*RGB* 颜色模型是最易量化的模型，它通过红色、绿色和蓝色的搭配来

精准地构造需要的颜色。在图像检索技术中，因为检索过程要符合人的主观意识，而人眼对于 RGB 模型不如 HSV 模型敏感，实际中人眼对于 HSV 颜色模型更加容易感知。

(1) HSV 颜色模型。HSV 模型即色调 H、饱和度 S 和亮度 V，此模型可以用三维坐标系统表示，如图 7-4 所示。

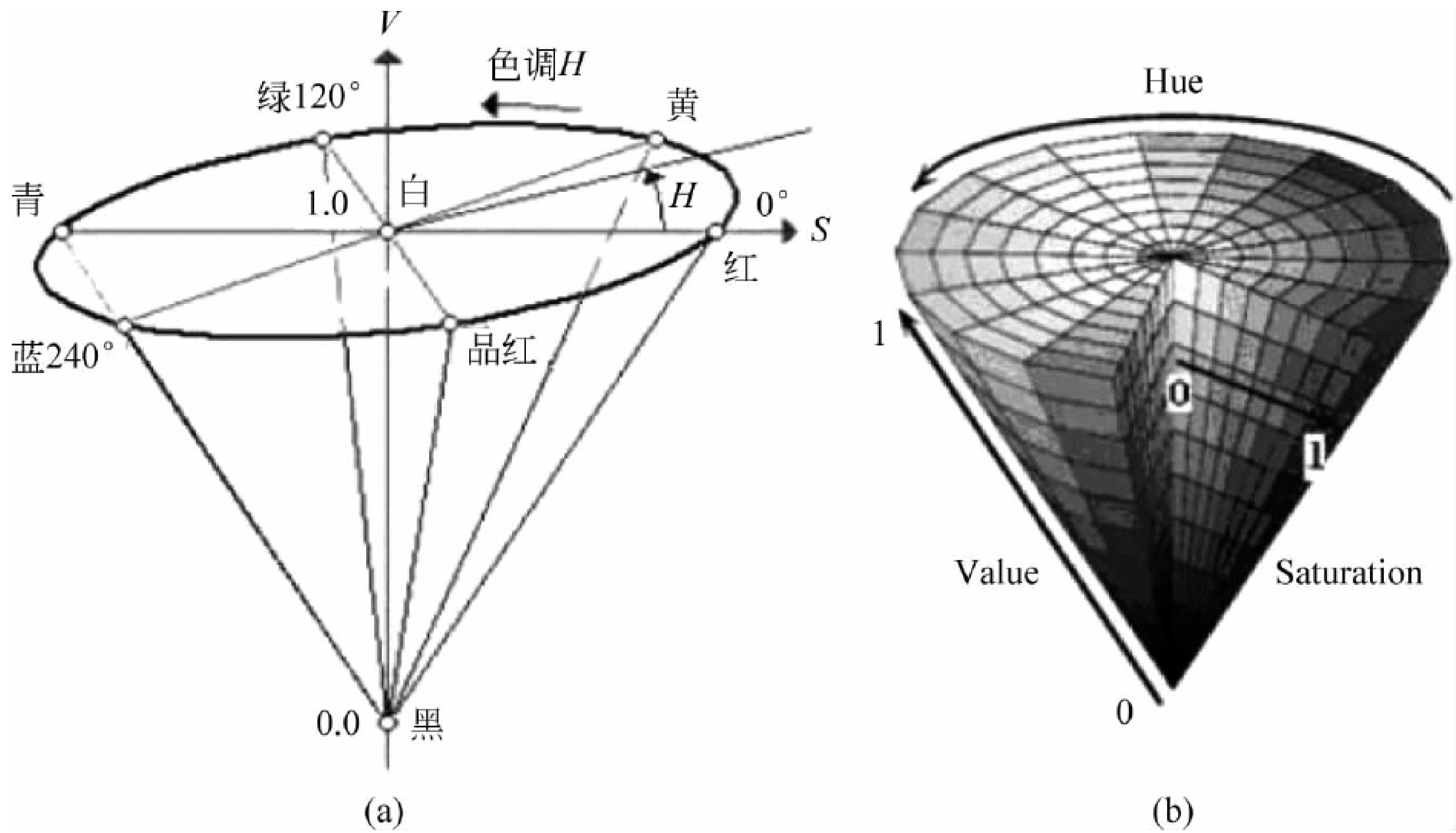


图 7-4　HSV 颜色模型图

首先将 RGB 量化为 HSV，再进行分维操作，即量化成若干个等级。实验表明，维数越多计量不一定越精准，维数与检索精度不成线性关系，当维数增加到一定数量时，增加维度来增加检索精度可能不明显甚至出现倒退。进行分维是必要的，比如同样的取景，一张是以晴天为背景的照片，一张是以阴天为背景的照片，在人眼中，这是相同的两张照片，但是由于色调、饱和度和亮度的细微差别，在计算机中，可能被认为是完全不同的两张照片，进行分维让计算机能像人眼一样忽略其中某些细微差别。将 HSV 空间（$h\in[0,360]$，$s\in[0,1]$，$v\in[0,1]$）非均匀量化为 32 类，这里进行简要描述如下：

If $v<0.2$　黑色 row＝0

Else if $v\geqslant0.2$ and $s<0.1$　根据 $v\in[0.2,1.0]$ 划分为三类：深灰[0.2,0.5)，浅灰[0.5,0.8)，白色[0.8,1.0]row＝1,2,3

Else $s\geqslant0.1$

将 H 非均等分成赤色[0,20)，橙色[20,45)，黄色[45,75)，绿色[75,165)，青色[165,200)，蓝色[200,270)，紫色[270,360]七个部分。

将 V 分为暗色[0.2,0.5)，明色[0.5,1.0)两个部分。

将 S 分为浅色[0.1,0.45)，浓色[0.45,1.0)两个部分，共 $7\times2\times2=28$ 种划分，row∈[4,32]且 row∈N^*。用上述方法将 HSV 色彩模型共分为 $1+3+28=32$ 种颜色。

(2) RGB 颜色模型。我们日常见到的最普遍的颜色模型就是 RGB 模型，它与人眼视觉结构密切相关，它是一个三维空间模型，三个坐标轴分别是 R(红)，G(绿)，B(蓝)轴，组成一个单位正方体，坐标轴的原点是黑色，离原点最远的顶点是白色，立方体与三个轴的焦点分别是紫色、蓝绿色、黄色。计算机中的数字图像一般是用 RGB 颜色模型来表示的，对于三个分量，单位由位(bit)来表示，范围是 0～255，RGB 模型的优点是方便计算机统计和存储，但 RGB 模型是颜色分布最不均匀的模型之一，难以用距离来衡量两种不同的颜色，不符合人眼的直观感知，也就是给你一组 RGB 数据，你很难想象它的实际颜色。

(3) YUV 颜色模型。YUV 颜色模型又称 YCrCb 模型，是欧洲电视系统所采用的颜色模型。Y 表示亮度信号，U、V 表示色度信号，这个模型也是根据人眼对颜色分辨程度来划分的，从 RGB 模型变换为 YUV 模型是线性变换，公式描述如下：

$$\begin{bmatrix} Y \\ U \\ V \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -00147 & -0.289 & 0.436 \\ 0.615 & -0.515 & -0.100 \end{bmatrix} * \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{7-1}$$

2. 颜色特征提取

(1) 颜色直方图。颜色直方图的优点在于它能简单描述一幅图像中颜色的全局分布，即不同色彩在整幅图像中所占的比例，特别适用于描述那些难以自动分割的图像和不需要考虑物体空间位置的图像。该方法的缺点在于它无法描述图像中颜色的局部分布及每种色彩所处的空间位置，即无法描述图像中的某一具体的对象或物体。颜色直方图最常用的颜色空间是 RGB 颜色空间和 HSV 颜色空间。颜色直方图特征匹配方法主要有直方图相交法、距离法、中心距法、参考颜色表法、累加颜色直方图法等。

颜色直方图的生成是对于图像进行顺序逐行完全扫描，记录每一种颜色在整个图像颜色集中出现的次数，得出其出现的频数。$(f_{xy})_{M\times N}$ 表示给定的图像，$M\times N$ 是这幅图像的分辨率，C 表示这幅图像的颜色集，f_{xy} 表示给定点 (x,y) 处的颜色值，则图像的色彩直方图公式如下：

$$h_c = \frac{1}{M\times N}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\delta(f_{ij}-c), \quad \forall c\in C \tag{7-2}$$

颜色直方图作为传统的色彩特征提取方法，很好地体现了色彩特征提取的优点：易提取，

易统计,对图像大小变换、旋转、平移不敏感。但色彩直方图缺点也是明显的:丢失色彩的空间信息,容易受到背景噪音影响,特征维数高。因此很多基于此算法的改进算法被提出,例如将色彩特征非等量化为 77 维向量,忽略了出现频度低以及人眼难以辨别的颜色,新的算法使得直方图对于背景噪音的敏感度降低了。

(2) 累加直方图。由于大多数的直方图非常稀疏且对噪音敏感,有学者采用直方图的累积法。其中,由于累加直方图体现了两种颜色在颜色轴上的距离与相似性之间的关系,所以累加直方图法在检索效率上优于一般直方图法。但累加直方图能体现这个优势的前提是:信号本身在特征分布轴上距离小的两点要比距离大的两点更相似。人的视觉特性对上述相关性条件在整个色度分布轴上并不成立,但在色度分布轴上的各个局部区间里能够满足,所以我们把色度沿分布轴分成若干个局部区间,而在各局部区间内分别应用累加直方图法。

累加直方图在具体检索时,先将色度轴分成六个不重叠的局部区间$[60k,60(k+1)]$,$k=0,1,\cdots,5$,然后分别计算每个局部区间的累加直方图。由于色度轴上各种颜色的分布实际上是连续过渡的,各颜色区之间并不存在截然的界限,区间改变为$[30+60k,(30+60(k+1))\bmod 360]$,$k=0,1,\cdots,5$,计算出这时每个局部区间的累加直方图。最后将这两次计算的累加直方图逐项相加取平均,作为最终的特征直方图用于检索。检索实验证明,这种局部累加直方图法在检索效率上要远远优于一般累加直方图法。

(3) 颜色矩和颜色集。除了颜色直方图外,在图像检索中颜色矩和颜色集也用于表示图像特征。

① 颜色矩。颜色矩(color moments)最早是由 Stricker 等人提出的,一幅图像的颜色信息通常分布在低阶矩中,所以在实际应用中,一般只用到一阶矩(mean)、二阶中心矩(vanriance)及三阶中心矩(skewness)。这三个低阶矩的表达式为

$$u_i = \frac{1}{n}\sum_{j=1}^{n} h_{ij} \tag{7-3}$$

$$\sigma_i = \left[\frac{1}{n}\sum_{j=1}^{n}(h_{ij}-u_i)^2\right]^{\frac{1}{2}} \tag{7-4}$$

$$\zeta_i = \left[\frac{1}{n}\sum_{j=1}^{n}(h_{ij}-u_i)^3\right]^{\frac{1}{3}} \tag{7-5}$$

h_{ij} 表示第 i 个颜色分量中灰度值为 j 的像素出现的频度,n 表示灰度级。以 *RGB* 颜色模型为例,这个颜色模型具有三个颜色分量,所以一般统计一幅图像颜色低阶矩时,一共有 $3\times3=9$ 个常用分量。颜色矩的优点是表达简洁,易于计算,缺点是检索效率低,在

实际的应用中,颜色矩一般作为辅助数据并联合其他图像特征一起进行检索工作。

② 颜色集方法。为了提高图像检索速度采用颜色集方法,首先将 RGB 颜色空间转换成视觉均衡的颜色空间(HSV),并将颜色空间量化成若干个 bin,然后运用颜色自动分割技术将图像分为若干个区域,每个区域用量化颜色空间的某个颜色分量来索引,从而将图像表达成一个二进制的颜色索引表。在图像匹配中,比较不同图像颜色集之间的距离和颜色区域的空间关系。因为,颜色集表达为二进制的特征向量,可以构造二分叉树来加快检索速度,对大规模的图像集合十分有利。

7.3.2 基于纹理特征的图像检索

纹理特征是一种不依赖于颜色或亮度的反映图像中同质现象的视觉特征。它是所有物体表面共有的内在特性,例如云彩、树木、砖、织物、动物皮肤等都有各自的纹理特征(例如门禁系统中指纹识别就是图像纹理特征应用的典型例子)。纹理特征包含了物体表面结构组织排列的重要信息以及它们与周围环境的联系。正因为如此,纹理特征在基于内容的图像检索中得到了广泛的应用,用户可以通过提交包含有某种纹理的图像来查找含有相似纹理的其他图像。

早在 20 世纪 70 年代,产生了共生矩阵(co-occurrence matrix)表示图像纹理特征的方法。该方法从数学角度研究了图像纹理中灰度级的空间依赖关系。它首先建立一个基于像素间方向性和距离的共生矩阵,然后从矩阵中提取有意义的统计量作为纹理特征。因为图像中相距($\Delta x,\Delta y$)的两个灰度像素同时出现的联合频率分布可以用灰度共生矩阵来表示。若将图像的灰度级定为 N 级,那么共生矩阵为 $N\times N$ 矩阵,可表示为 $M_{(\Delta x,\Delta y)}(x,y)$,其中位于($h,k$)的元素 m_{hk} 的值表示一个灰度为 h 而另一个灰度为 k 的两个相距为($\Delta x,\Delta y$)的像素对出现的次数。

对粗纹理的区域,其灰度共生矩阵中的 m_{hk} 值较集中于主对角线附近。因为对于粗纹理,像素对趋于具有相同的灰度。而对于细纹理的区域,其灰度共生矩阵中的 m_{hk} 值则散布在各处。由此可见用灰度共生矩阵的各种统计量可作为纹理特性的度量。通常利用以下四个特征量表示图像的纹理特征。

1. 反差(或称为主对角线的惯性矩)

$$\mathrm{CON} = \sum_{h} \sum_{k} (m_{hk})^2 \tag{7-6}$$

对于粗纹理,由于 m_{hk} 的数值较集中于主对角线附近,此时 h-k 的值较小,所以相应的 CON 值也较小。相反,对于细纹理则相应的 CON 值较大。

2. 能量(或称为角二阶矩)

$$\mathrm{ASM}=\sum_{k}\sum_{k}(m_{hk})^{2} \tag{7-7}$$

这是一种对图像灰度分布均匀性的度量,当 m_{hk} 的数值分布较集中于主对角线附近时,其相应的 ASM 值较大;反之,ASM 值则较小。

3. 熵

$$\mathrm{ENT}=\sum_{k}\sum_{k}m_{hk}\log m_{hk} \tag{7-8}$$

当灰度共生矩阵中各 m_{hk} 数值相差不大且较分散时,ENT 值较大;反之,若 m_{hk} 的数值较集中时,ENT 值较小。

4. 相关

$$\mathrm{COR}=\left[\sum_{k}\sum_{k}hkm_{hk}-\mu_{x}\mu_{y}\right]\Big/\delta_{x}\delta_{y} \tag{7-9}$$

其中 $\mu_x,\mu_y,\delta_x,\delta_y$ 分别为 m_x,m_y 的均值和标准差,$m_x=\sum_{k}m_{hk}$ 是矩阵 $\boldsymbol{M}$ 中每列元素之和;$m_y=\sum_{h}m_{hk}$ 是矩阵 $\boldsymbol{M}$ 中每行元素之和。相关量是用来描述矩阵中行或列元素之间相似程度的,它是灰度线性关系的度量。

在纹理特征的提取中,我们先把图像的亮度分量图分成 64 个灰度级,并构造四个方向的共生矩阵即 $\boldsymbol{M}_{(1,0)},\boldsymbol{M}_{(0,1)},\boldsymbol{M}_{(1,1)},M_{(1,-1)}$,然后分别计算四个共生矩阵的上述四个纹理参数,最后以各参数的均值和标准差即 $\mu_{\mathrm{CON}},\delta_{\mathrm{CON}},\mu_{\mathrm{ASM}},\delta_{\mathrm{ASM}},\mu_{\mathrm{ENT}},\delta_{\mathrm{ENT}},\mu_{\mathrm{COR}},\delta_{\mathrm{COR}}$ 作为纹理特征向量中的各个分量。由于以上八个分量的物理意义和取值范围不同,需对它们进行内部归一化。这样在计算相似距离时,可使各分量具有相同权重。高斯归一化方法是一种较好的归一化方法,其特点是少量的超大或超小的元素值对整个归一化后的元素值分布影响不大,具体方法如下。

一个 N 维的特征向量可记为:$\boldsymbol{F}=[f_1,f_2,\cdots,f_N]$。如用 $I_1,I_2,\cdots,I_M$ 代表图像库中的图像,则对其中任一幅图像 I_i,其相应的特征向量为 $\boldsymbol{F_i}=[f_{i,1},f_{i,2},\cdots,f_{i,N}]$。假设特征分量值系列$[f_{1,j},f_{2,j},f_{i,j},\cdots,f_{M,j}]$符合高斯分布,计算出其均值 m_j 和标准差 δ_j,然后利用下式可将 $f_{i,j}$ 归一化至$[-1,1]$区间,公式如下:

$$f_{i,j}^{(N)}=\frac{f_{i,j}-m_j}{\delta_j} \tag{7-10}$$

根据上式归一化后,各个 $f_{i,j}$ 均转变成具有 $N(0,1)$分布的 $f_{i,j}^{(N)}$。如果利用 δ_j 进行归一化,则 $f_{i,j}^{(N)}$ 的值落在$[-1,1]$区间的概率可达 99%。实际应用中,将$[-1,1]$区间外的 $f_{i,j}$ 值设为-1 或 1,以保证所有 $f_{i,j}$ 的值均落在$[-1,1]$区间。

基于人类对纹理的视觉感知的心理学研究，可以从另一个角度提出纹理特征的表达：Tamura 纹理特征。Tamura 纹理特征的六个分量对应于心理学角度上纹理特征的六种属性，分别是粗糙度（coarseness）、对比度（contrast）、方向度（directionality）、线像度（1inelikeness）、规整度（regularity）和粗略度（roughness）。

Tamura 纹理特征和共生矩阵的一个主要不同是 Tamura 纹理特征中的所有纹理属性有视觉意义，而共生矩阵中的一些纹理属性却没有（如熵）。这一特征使得 Tamura 纹理特征在图像检索中很受欢迎。

20 世纪 90 年代初，由于小波变换的出现及其理论框架的建立，许多研究人员开始研究在纹理表示时用小波变换。例如，用子带小波中提取的统计量作为纹理特征。这种方法检索纹理图像时准确率超过 90%。为了提取中带特征，可以采用树结构小波变换来进一步提高分类的准确率。此外，小波变换也常常与其他技术结合以获得更好的效果，例如正交和双正交小波变换、树结构小波变换以及 Gabor 小波变换。

7.3.3 基于形状特征的图像检索

图像内容的形状是揭示物体的本质特征之一，可以针对面积（可用像素点的个数计算）、环形性（即周长×周长/面积，周长也用像素点的个数表示）、主轴方向、偏心率、圆形率、连通性、正切角等形状特征进行匹配。通常来说，图形内容的形状特征有两种表示方法：一种是轮廓特征，一种是区域特征。前者只用到物体的外边界，而后者则关系到整个形状区域。这两类形状特征的最典型方法分别是傅里叶描述符和形状无关矩。傅里叶形状描述符（Fourier shape descriptors）的基本思想是用物体边界的傅里叶变换作为其形状描述。

1. 形状特征提取的一般几何原理

1）矩形度

矩形度反映目标对其外接矩形的充满程度，用目标的面积与其最小外接矩形的面积之比来描述，即

$$R=\frac{A_{O}}{A_{MER}} \tag{7-11}$$

式中，A_O 是该目标的面积，而 A_{MER} 是 MER 的面积。R 的值为 0～1，当目标为矩形时，R 取得最大值 1.0；圆形目标的 R 取值为 $\pi/4$；细长的、弯曲的目标的 R 取值变小。

另外一个与形状有关的特征是长宽比 r：

$$r=\frac{W_{MER}}{L_{MER}} \tag{7-12}$$

r 即为 MER 宽与长的比值，利用 r 可以将细长的目标与圆形或方形的目标区分开来。

2）圆形度

（1）致密度 C。度量圆形度最常用的是致密度，即周长（P）的平方与面积（A）的比。

$$C = \frac{P^2}{A} \tag{7-13}$$

（2）边界能量 E。边界能量是圆形度的另一个指标。假定目标的周长为 P，用变量 p 表示边界上的点到某一起始点的距离。边界上任一点都有一个瞬时曲率半径 $r(p)$，该点与边界相切圆的半径 p 点的曲率函数是

$$K(p) = \frac{1}{r(p)} \tag{7-14}$$

函数 $K(p)$ 是周期为 P 的周期函数。可用下式计算单位边界长度的平均能量：

$$E = \frac{1}{P}\int_0^p |K(p)^2| \mathrm{d}p \tag{7-15}$$

在面积相同的条件下，圆具有最小边界能量 $E_0 = (2\pi/P)2 = (1/R)2$，其中 R 为圆的半径。曲率可以很容易地由链码算出，因而边界能量也可方便算出。瞬时曲率半径 $r(p)$ 与边界能量示意图见图 7-5。

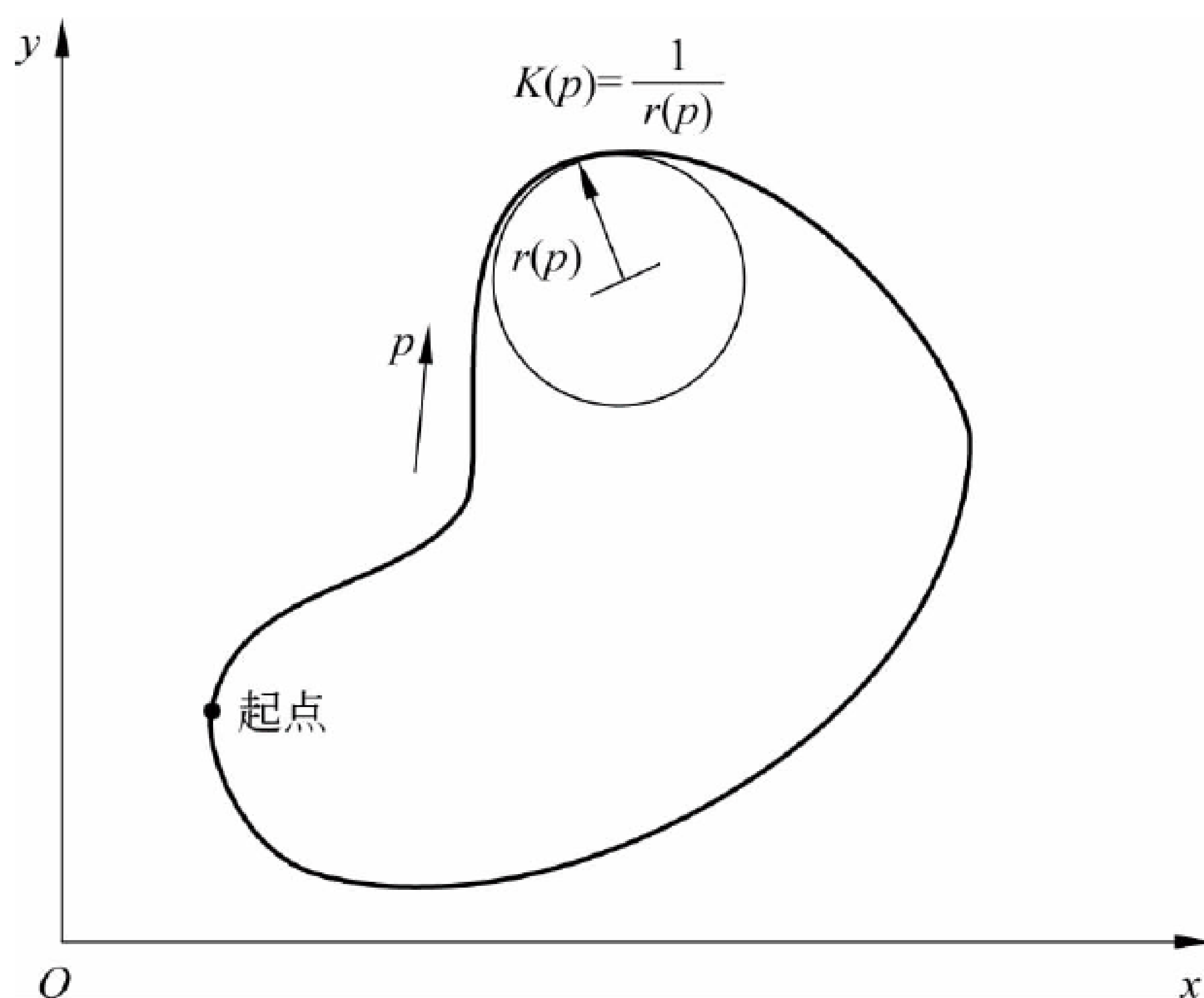


图 7-5　瞬时曲率半径 $r(p)$ 与边界能量示意图

（3）圆形性。圆形性（circularity）C 是一个用区域 R 的所有边界点定义的特征量，即

$$C=\frac{\mu_R}{\delta_R} \tag{7-16}$$

式中，μ_R 是从区域重心到边界点的平均距离，δ_R 是从区域重心到边界点的距离均方差：

$$\mu_R=\frac{1}{K}\sum_{k=0}^{K-1}||(x_k,y_k)-(\bar{x},\bar{y})|| \tag{7-17}$$

$$\delta_R=\frac{1}{K}\sum_{k=0}^{K-1}[||(x_k,y_k)-(\bar{x},\bar{y})||-\mu_R]^2 \tag{7-18}$$

当区域 R 趋向圆形时，特征量 C 是单调递增且趋向无穷的，它不受区域平移、旋转和尺度变化的影响，可以推广用于描述三维图像目标。

（4）面积与平均距离平方的比值。圆形度的第四个指标利用了从边界上的点到目标内部某点的平均距离，即

$$\bar{d}=\frac{1}{N}\sum_{i=1}^{N}x_i \tag{7-19}$$

式中，x_i 是从具有 N 个点的目标中的第 i 个点到与其最近的边界点的距离。相应的形状度量为

$$g=\frac{A}{d}=\frac{N^3}{\sum_{i=1}^{N}x_i} \tag{7-20}$$

3）球状性

球状性(sphericity)S，既可以描述二维目标也可以描述三维目标，其定义为

$$S=\frac{r_i}{r_c} \tag{7-21}$$

在二维情况下，r_i 代表区域内切圆(inscribed circle)的半径，而 r_c 代表区域外接圆(circumscribed circle)的半径，两个圆的圆心都在区域的重心上。

当区域为圆时，球状性的值 S 达到最大值1.0，而当区域为其他形状时，则有 $S<1.0$。S 不受区域平移、旋转和尺度变化的影响。图像形状的球状性定义见图 7-6。

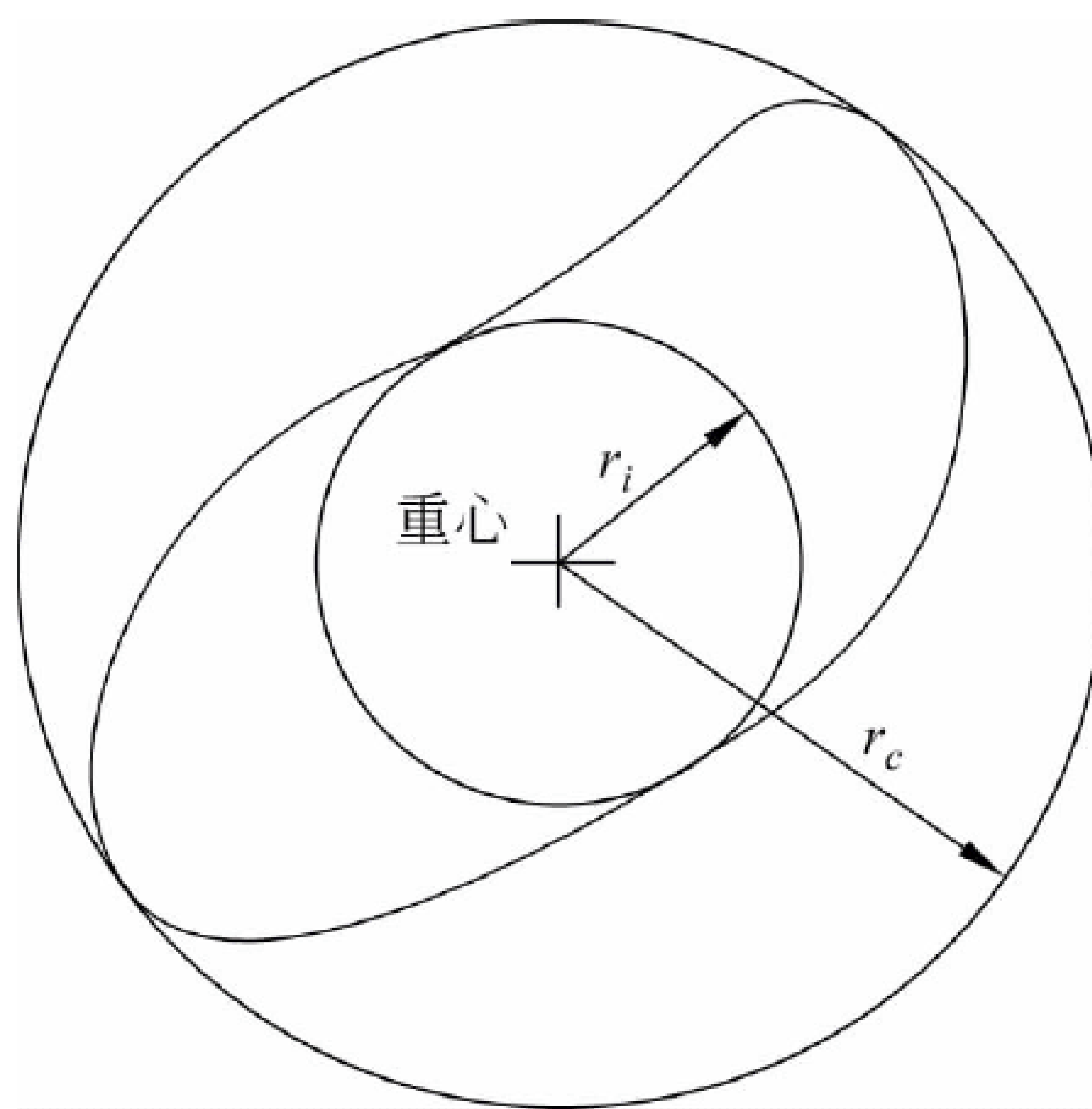


图 7-6　图像形状的球状性定义

4）不变矩

（1）矩的定义。对于二元有界函数 $f(x,y)$，它的$(j+k)$阶矩为

$$M_{jk}=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}x^j y^k f(x,y)\mathrm{d}x\mathrm{d}y \quad j,k=0,1,2,\cdots \tag{7-22}$$

由于 j 和 k 可取所有的非负整数值,因此形成了一个矩的无限集。而且,这个集合完全可以确定函数 $f(x,y)$本身。换句话说,集合$\{M_{jk}\}$对于函数 $f(x,y)$是唯一的,也只有 $f(x,y)$才具有这种特定的矩集。

(2) 质心坐标与中心矩。当 $j=1,k=0$ 时,M_{10}对二值图像来讲就是目标上所有点的 x 坐标的总和,类似地,M_{01}就是目标上所有点的 y 坐标的总和,所以

$$\bar{x}=\frac{M_{10}}{M_{00}},\quad \bar{y}=\frac{M_{01}}{M_{00}} \tag{7-23}$$

就是二值图像中一个目标的质心的坐标。

为了获得矩的不变特征,往往采用中心矩以及归一化的中心矩。中心矩的定义为

$$M'_{jk}=\sum_{x=1}^{N}\sum_{y=1}^{M}(x-\bar{x})^j(y-\bar{y})^k f(x,y) \tag{7-24}$$

(3) 主轴。使二阶中心矩从 μ_{11}变得最小的旋转角 θ 可以由下式得出:

$$\tan 2\theta=\frac{2\mu_{11}}{\mu_{20}-\mu_{02}} \tag{7-25}$$

将 x、y 轴分别旋转 θ 角得坐标轴 x'、y',称为该目标的主轴。上式中在 θ 为 60°时的不确定性

$$\mu_{20}<\mu_{02},\quad \mu_{30}>0$$

可以通过条件限定中心矩来解决。如果目标在计算矩之前旋转 θ 角,或相对 x'与 y'轴计算矩,那么矩具有旋转不变性。

(4) 不变矩。相对于主轴计算并用面积归一化的中心矩,在目标放大、平移、旋转时保持不变。只有三阶或更高阶的矩经过这样的规一化后才能保持不变性。对于 $j+k=2,3,4,\cdots$的高阶矩,可以定义归一化的中心矩为

$$\mu_{jk}=\frac{M'_{jk}}{(M_{00})^r},\quad r=\left(\frac{j+k}{2}+1\right) \tag{7-26}$$

利用归一化的中心矩,可以获得六个不变矩组合,这些组合对于平移、旋转、尺度等变换都是不变的,不变矩组合如下:

$$\phi_1=\mu_{20}+\mu_{02}$$
$$\phi_2=(\mu_{20}-\mu_{02})^2+4\mu_{11}^2$$
$$\phi_3=(\mu_{30}-3\mu_{12})^2+(\mu_{03}-3\mu_{21})^2$$

$$\phi_4 = (\mu_{30} + \mu_{12})^2 + (\mu_{03} + \mu_{21})^2$$

$$\phi_5 = (\mu_{30} - 3\mu_{12})(\mu_{03} + \mu_{12}) \times [(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2] + (\mu_{03} - 3\mu_{21})(\mu_{30} + \mu_{21}) \times [(\mu_{03} + \mu_{21})^2 - 3(\mu_{12} + \mu_{30})^2]$$

$$\phi_6 = (\mu_{20} - \mu_{02})[(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2] + 4\mu_{11}(\mu_{30} + \mu_{21})(\mu_{03} + \mu_{21})$$

不变矩及其组合具备了良好形状特征应具有的某些性质，已经用于印刷体字符的识别、飞机形状区分、景物匹配和染色体分析中，但它们并不能确保在任意情况下都具有这些性质。一个目标形体的唯一性体现在一个矩的无限集中，因此，要区别相似的形体需要一个很大的特征集。这样所产生的高维分类器对噪声和类内变化十分敏感。在某些情况下，几个阶数相对较低的矩可以反映一个目标的显著形状特征。图像的形状特征提取一般是针对图像的一定区域展开的，图像的各个区域形状特征组合为图像的整体形状特征。区域形状特征的提取有三类方法：区域内部（包括空间域和变换域）形状特征提取方法、区域外部（包括空间域和变换域）形状特征提取方法和利用图像层次型数据结构提取形状特征方法。

5）偏心率

偏心率（eccentricity）E，也可叫伸长度（elongation），它在一定程度上描述了区域的紧凑性。偏心率 E 有多种计算公式，一种常用的简单方法是区域主轴（长轴）长度（A）与辅轴（短轴）长度（B）的比值，如图 7-7 所示。

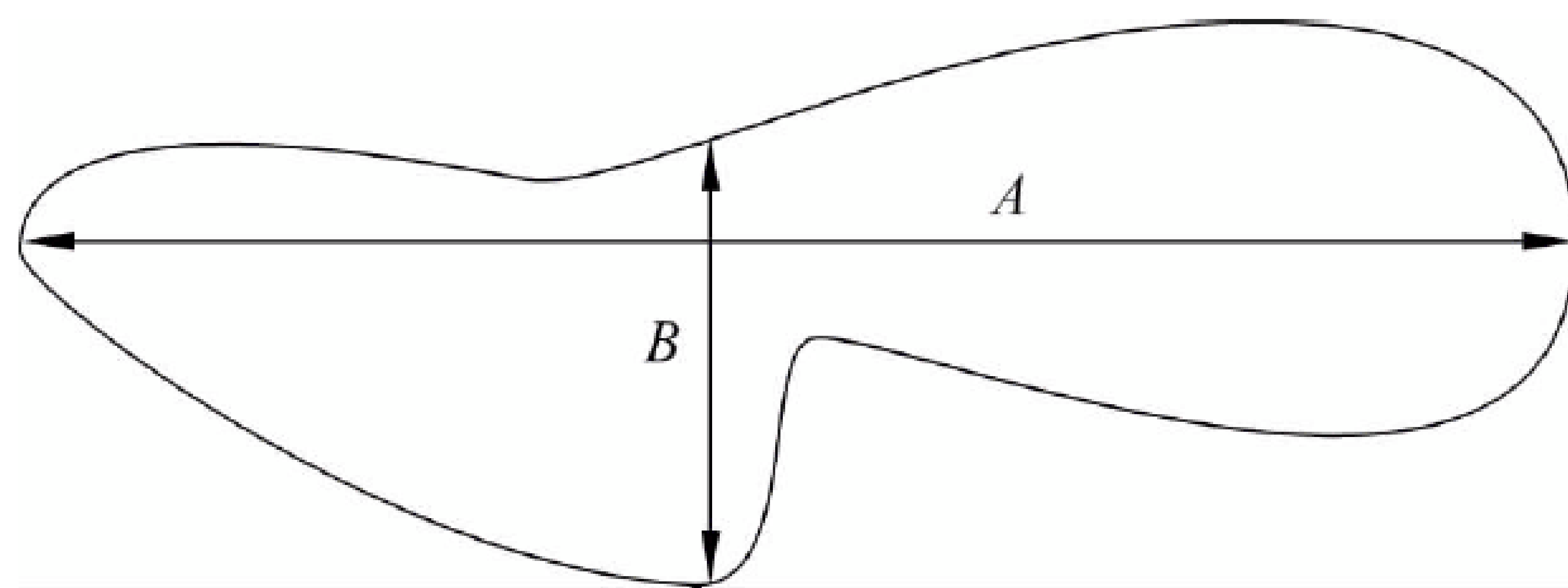


图 7-7　图像形状的偏心率示意图

在图 7-7 中，主轴与辅轴相互垂直，且其长度是两方向的最大值，不过这样的计算受目标形状和噪声的影响比较大。另一种方法是计算惯性主轴比，它基于边界线上的点或整个区域来计算向量。计算任意点集偏心度的近似公式，步骤如下：

第一步，计算平均向量，公式如下：

$$x_0 = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad y_0 = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{7-27}$$

第二步，计算 $j+k$ 阶中心矩，公式如下：

$$M_{jk}=\sum_{i=1}^{N}\sum_{i=1}^{N}(x_i-x_0)^j(y_i-y_0)^k \tag{7-28}$$

第三步，计算方向角，公式如下：

$$\theta=\frac{1}{2}\tan^{-1}\left(\frac{2M_{11}}{M_{20}-M_{02}}\right)+N\left(\frac{\pi}{2}\right) \tag{7-29}$$

第四步，计算偏心度的近似值，公式如下：

$$E=\frac{(M_{20}-M_{02})^2+4M_{11}}{A} \tag{7-30}$$

2. 图像形状特征提取的一般描述

1) 边界链码

链码是对图像边界点的一种编码表示方法，其特点是利用一系列具有特定长度和方向的相连直线段来表示目标的边界。因为每个线段的长度固定而方向数目有限，所以只有边界的起点需要用绝对坐标表示，其余点都可只用接续方向来代表偏移量。由于表示一个方向的比特数比表示一个坐标值所需比特数少，而且对每一个点又只需一个方向数就可以代替两个坐标值，因此链码表达可大大减少边界表示所需的数据量。

图像一般是按固定间距的网格采集的(点阵图像)，因此最简单的链码是跟踪边界并赋给每两个相邻像素连线为一个方向值。常用方法的有 4 方向链码和 8 方向链码，其方向定义分别如图 7-8(a)、图 7-8(b)所示，其中图 7-8(a)为 4 方向链码；图 7-8(b)为 8 方向链码；图 7-8(c)为边界编码图形。它们的共同特点是直线段的长度固定，方向数有限。

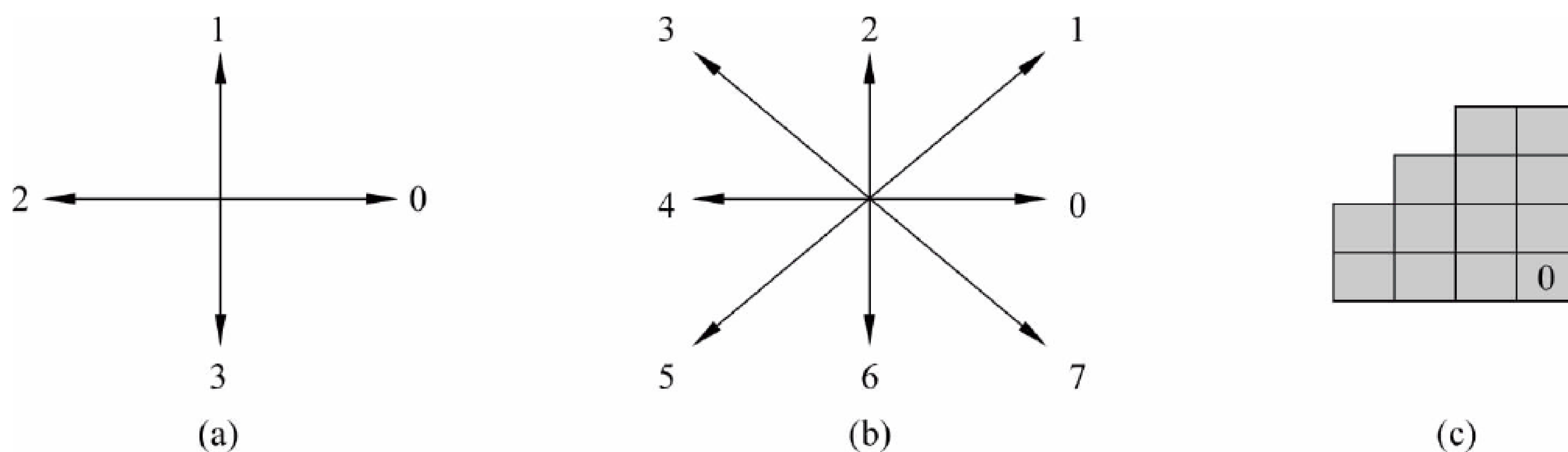


图 7-8　码值与方向对应关系图

对图 7-8(c)所示边界，若设起始点 O 的坐标为(5,5)，则分别用如下 4 方向和 8 方向链码表示区域边界。

4 方向链码：(5,5)1 1 1 2 3 2 3 2 3 0 0 ；

8 方向链码：(5,5)2 2 2 4 5 5 6 0 0。

图像特征实际提取中，直接对图像分割所得的目标边界进行编码有可能出现两个问题：一是码串比较长，二是噪声等干扰会导致小的边界变化从而使链码发生与目标整体形状无关的较大变动。常用的改进方法是对原边界以较大的网格重新采样，并把与原边界点最接近的大网格点定为新的边界点。这种方法也可用于消除目标尺度变化链码的影响。

使用链码时，起点的选择是很关键的。对同一个边界，如用不同的边界点作为链码的起点，得到的链码则是不同的。为解决这个问题可把链码归一化，给定一个从任意点开始产生的链码，可把它看做一个由各方向数构成的自然数。首先，将这些方向数按照一个方向循环，以使它们所构成的自然数的值最小；然后，将这样转换后所对应的链码起点作为这个边界的归一化链码的起点。

2）一阶差分链码

用链码表示给定目标的边界时，如果目标平移，链码不会发生变化，而如果目标旋转则链码会发生变化。为解决这个问题，可利用链码的一阶差分来重新构造一个表示原链码各段之间方向变化的新序列，这相当于把链码进行旋转归一化。差分可用相邻两个方向数按反方向相减（后一个减去前一个）得到。见图 7-9。

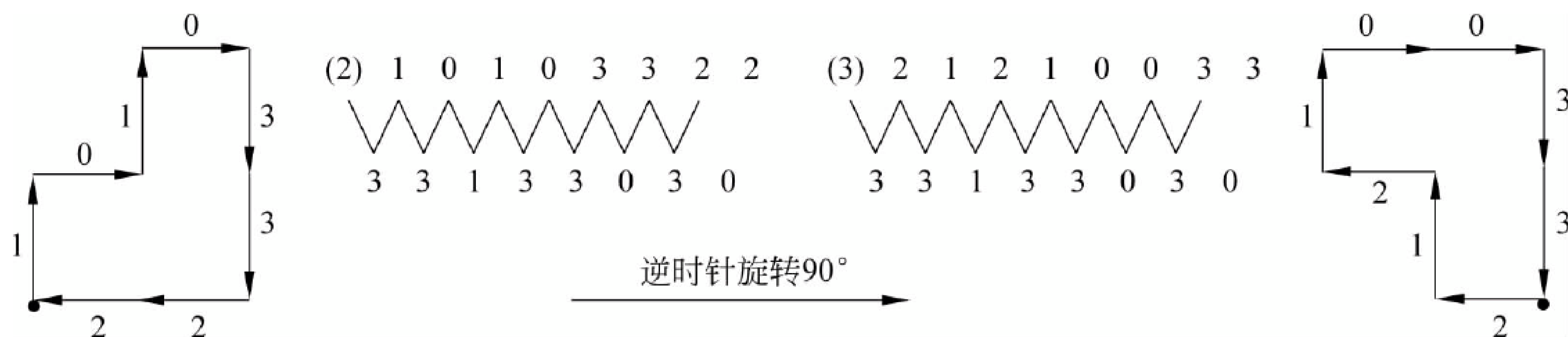


图 7-9　利用一阶差分对链码旋转归一化

如图 7-9 所示，上面一行为原链码（括号中为最右一个方向数循环到左边），下面一行为上面一行的数两两相减得到的差分码。左边的目标在逆时针旋转 90°后成为右边的形状，可见，原链码发生了变化，但差分码并没有变化。

3）傅里叶描述

对边界的离散傅里叶变换表达，可以作为定量描述边界形状的基础。采用傅里叶描述的一个优点是将二维问题简化为一维问题。即将 x-y 平面中的曲线段转化为一维函数 $f(r)$（在 r-$f(r)$ 平面上），也可将 x-y 平面中的曲线段转化为复平面上的一个序列。具体

就是将 x-y 平面与复平面 u-v 重合，其中，实部 u 轴与 x 轴重合，虚部 v 轴与 y 轴重合。这样可用复数 $u+jv$ 的形式来表示给定边界上的每个点(x,y)。这两种表示在本质上是一致的，是点点对应的(见图 7-10)。

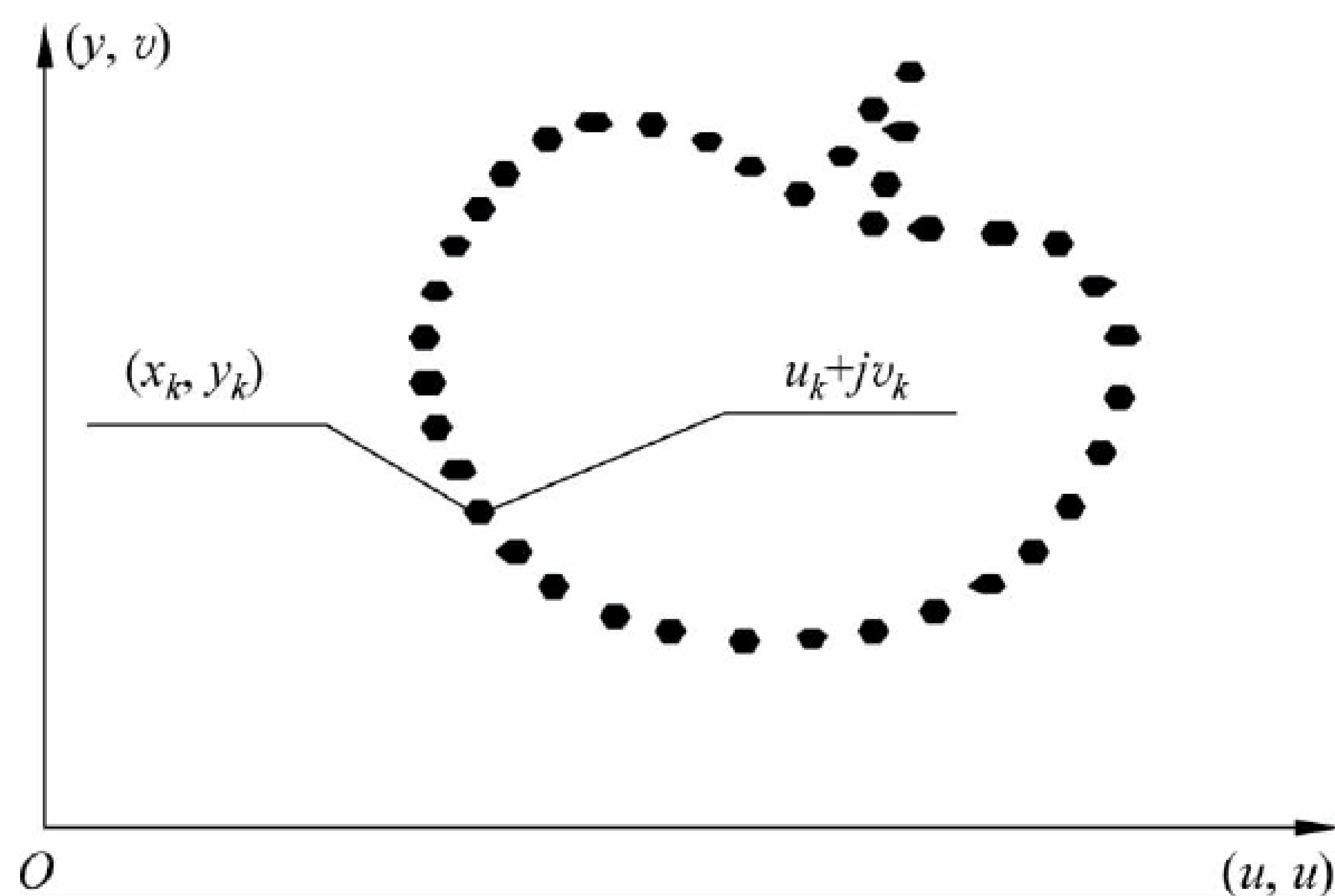


图 7-10　图像边界点的两种表示方法

如果一个由 N 个点组成的封闭边界，从任一点开始绕边界一周就得到一个复数序列，即

$$s(k)=u(k)+jv(k)\quad k=0,1,\cdots,N-1 \tag{7-31}$$

$s(k)$的离散傅里叶变换是

$$S(\omega)=\frac{1}{N}\sum_{k=0}^{N-1}s(k)\exp\left(-j\,\frac{2\pi\omega k}{N}\right),\quad \omega=0,1,\cdots,N-1 \tag{7-32}$$

$S(\omega)$可称为边界的傅里叶描述，它的傅里叶逆变换是

$$S(k)=\frac{1}{N}\sum_{\omega=0}^{N-1}S(\omega)\exp\left(-j\,\frac{2\pi\omega k}{N}\right),\quad k=0,1,\cdots,N-1 \tag{7-33}$$

可见，离散傅里叶变换是个可逆线性变换，在变换过程中信息没有任何增减，但这为有选择地描述边界提供了方便。只取 $S(\omega)$的前 M 个系数即可得到 $s(k)$的一个近似：

$$\bar{s}(k)=\frac{1}{N}\sum_{\omega=0}^{M-1}S(\omega)\exp\left(-j\,\frac{2\pi\omega k}{N}\right),\quad k=0,1,\cdots,N-1 \tag{7-34}$$

上式中 k 的范围不变，即在近似边界上的点数不变，但 ω 的范围缩小了，即为重建边界点所用的频率项少了。傅里叶变换的高频分量对应一些图像形状细节而低频分量对应图像总体形状，因此用一些低频分量的傅里叶系数足以近似描述边界形状。

3. 图像形状特征提取的研究进展

近年来，在形状表示和匹配方面的工作包括有限元方法(FEM)、旋转函数和小波描

述符。FEM定义一个稳定性矩阵来描述物体上的每一个点与其他点之间的联系。这个稳定性矩阵的特征向量被称为特征空间的模合基。所有的形状都首先被映射到该空间并通过特征值计算相似度。类似于傅里叶描述的思路,例如把旋转函数用来比较凹面和凸面多边形的相似性。用小波变换来描述物体形状。它几乎包含了符合要求的所有性质,如多分辨率表示、不变性、单一性、稳定性和空间位置等。就形状匹配算法而言,Chamfer匹配方法有较多成果;Chamfer匹配技术,该方法能够以线性的时间复杂度比较两个的形状块集合;分层Chamfer匹配算法可以加快匹配的速度,这种匹配算法可以在不同的精确层次上进行,逐步从粗糙到精确。

另外几何矩方法(基于区域)和傅里叶描述符(基于边缘)通过一种简单的线性变换联系起来。综合表示法基于某些特征(链编码、傅里叶描述符、UNL傅里叶描述符)的边缘表示法的效果,基于另一些特征(矩无关性、Zernike矩、pseudo-Zernike矩)的区域表示法的效果以及综合表示法(矩无关性和UNL傅里叶描述符、矩无关性和傅里叶描述符)。实验表明,综合表示法要优于单一的描述。

除了二维形状表示法外,三维形状特征表示的方法也很多。傅里叶描述符的标准方法,它包含了所有形状信息而且计算高效,利用傅里叶描述符的良好插补能力来有效地表示三维空间中的形状。也有兼顾结构和统计方法的局部形状分析算法来表示三维形状特征。此外,用代数无关矩来同时表示二维空间的形状特征和三维空间的形状特征,这大大地减少了形状匹配的计算量。

7.3.4 基于空间特征的图像检索

上述的颜色、纹理和形状等多种特征反映的都是图像的整体特征,而无法体现图像中所包含的对象或物体。事实上,图像中对象所在的位置和对象之间的空间关系同样是图像检索中非常重要的特征。打个比方,蓝色的天空和蔚蓝的海洋在颜色直方图上是非常接近而难以辨别的。但如果我们指明是"处于图像上半部分的蓝色区域",则一般来说就可以区分天空和海洋。由此可见,包含空间关系的图像特征对图像检索有很大帮助。

图像空间关系特征是指图像中分割出来的多个目标之间相互的空间位置或相对方向关系,这些关系也可分为连接/邻接关系、交叠/重叠关系和包含/包容关系等。通常空间位置信息可以分为两类:相对空间位置信息和绝对空间位置信息。前一种关系强调的是目标之间的相对情况,如上下左右关系等,后一种关系强调的是目标之间的距离大小以及方位。显而易见,由绝对空间位置可推出相对空间位置,但表达相对空间位置信息比较简单。

空间关系特征的使用可加强对图像内容的描述区分能力,但空间关系特征常对图像或目标的旋转、反转、尺度变化等比较敏感。另外,实际应用中,仅仅利用空间信息往往是不够的,不能有效准确地表达场景信息。为了进行准确的图像检索,除了使用空间关系特征外,还需要其他特征来配合。

图像空间关系特征提取方法可分为两类：一类方法是首先对图像进行自动分割,划分出其中所含的对象或颜色区域,然后根据这些区域对图像索引;另一类方法则简单地将图像均匀划分成若干规则子块,对每个图像子块提取特征建立索引。基于图像分割方法中的图像空间关系特征主要包括二维符号串、空间四叉树和符号图像;基于图像子块的方法将图像预先等分成若干子块,然后分别提取每个子块的各种特征。

7.3.5　单个特征图像检索的不足

基于图像颜色特征的索引存在的主要问题是人对颜色特征的视觉感知方面考虑得仍然不够,虽然目前大多数基于颜色特征的图像检索采用了和人对颜色感知相一致的 HIS 颜色空间,但关于两种颜色之间的相似度的定义和视觉上人对相似颜色的判定仍然有一定的差距。从颜色特征的表示来看,各种形式的颜色直方图是最常用的表示方法。从颜色特征的相似形提取角度来检索两幅图像,一般指定相同并采用几十到几百维的高维直方图,实际上人对两图像画面的颜色的相似性判定主要考虑少数几种主要的颜色。不同的图像有不同的颜色集,对包含不同颜色集的两图像之间的相似性判定仍然需要进一步研究。

基于图像纹理特征的索引目前存在的主要问题是各种方法所选择的纹理特征集依赖于具体的纹理图像,往往是一种方法所选择的纹理特征集对表达一个纹理图像数据库比较有效,但对另一个纹理图像数据库来说就不一定管用。对于不同的纹理图像数据库如何进行纹理特征集的自动匹配运算仍需要进一步研究,也就是基于图像纹理特征的图像检索技术的通用性研究依然是个难点问题。

对于形状特征的图像检索,形状边界的自动提取一直是困扰图像处理领域多年的难题。形状特征提取是一件非常繁重的工作,对于大批量图像数据,此问题将显得更为突出。各种形状特征表达方法对形状信息的丢失非常严重;只有少量的形状特征表达方法和形状的几何变换无关。另外形状度量方法仍不具有很好的形状区分能力,不能有效表达形状之间的相似性。研究形状特征检索仍是基于内容检索中较具有挑战性的研究课题。

基于空间关系的索引存在的主要问题是如何保证各种空间关系与图像的旋转无关,如何实现空间特征的相似度量从定性到定量的转变,目前仍没有很好的研究成果。在图

像多重特征的相关反馈检索中,由于不同的特征其度量空间是不一样的,如何将这些距离转变为图像之间的相似度量空间并能准确地表示人对图像之间的相似性认识,是非常难的一件事情。总而言之,采用单一的图像特征向量对图像数据库进行查询不能很好地解决查询中准确率和查询效率之间的矛盾,如果采用高维数的特征向量又会降低查询的效率,采用低维数的特征向量会降低查询的准确率,因此可综合利用多特征进行图像检索。

7.4 基于多特征的图像检索

鉴于利用图像单个特征检索的缺点,可以综合利用图像的颜色、纹理、形状和空间特征的方法,计算特征提取向量。用户可以根据需要调整各个特征之间的权重关系,以便满足不同应用情况的查询。

7.4.1 综合颜色和形状特征的图像检索

颜色和形状是图像重要的特征之一,而颜色直方图没有考虑所含对象的形状特征,形状特征没有完善的数学模型,为了弥补二者的不足,我们可以通过结合颜色直方图的相似度和边界方向直方图的相似度进行检索。设 m 为查询图像,n 为数据库中的图像集合,D_y 代表基于颜色直方图的相似度,D_b 代表基于边界方向直方图的相似度。则两幅图像间的相似性如下计算:

$$D_{(m,n)} = \frac{\omega_y D_y + \omega_b D_b}{\omega_y + \omega_b} \tag{7-35}$$

通过实验验证,综合颜色和形状特征比使用单个特征确实提高了检索正确率。通过二者结合,不仅完全克服了叠加噪声的影响,而且提高了旋转变化时的稳定性。两幅不同的图像有可能会有相似的颜色直方图或边界直方图,但同时具有相似的两种直方图的概率较小,即综合检索可以减少误匹配,从而提高检索的精确度和准确率。

7.4.2 综合形状和空间特征的图像检索

目前,颜色的空间索引技术有两种:基于图像空间的固定划分方法和基于像素颜色的空间相关性的聚类方法。综合形状特征和空间位置关系特征可以较好地处理一些二值图像。由于二值属于人工图像,例如二值商标图像,部分二值图像是由一些边界分明的几何形状体构成的,因此可把一些二值图像看做是由一些具有显著形状特征的区域构成的

集合体,对这些集合体首先利用矩形特征进行形状的相似性度量,然后利用投影分类的方法匹配空间位置关系。该方法既考虑了二值图像内部各组成部分的形状特征,又兼顾了它们之间的空间位置关系,将整个检索过程分为初级检索与高级检索反馈求精两个阶段。由于该方法保证了整体与局部的一致性,因此具有良好的检索精度,与只利用图像的形状特征进行检索的实验结果相比,其检索结果更加符合人的视觉感知特性。

7.4.3　综合形状和纹理特征的图像检索

纹理特征是一种统计特征,具有旋转不变性,并具有较强的抗噪音能力。由于纹理不能单纯地由颜色或密度得到,它不能反映出事物的本质属性,受图像的分辨率影响很大,易受到光照、反射的影响。图像的形状信息不随图像颜色的变化而变化,是物体稳定的特征。利用形状特征进行检索可提高检索的准确性和效率。但是基于形状的检索法缺乏比较完善的数学模型,目标物体发生变形时检索结果不可靠,全面的描述目标形状对计算和存储有较大的要求。将形状特征和图像纹理特征相结合,同时利用半自动的图像分割技术提取图像边缘区域。在检索过程中,假设使用形状特征进行检索和纹理特征的排序位置分别为 r_1 和 r_2,则综合特征的排序位置为$(r_1+r_2)/2$。通过实验,这种方法的查全率和查准率较使用单一特征要高。

7.4.4　综合颜色、形状和空间的图像检索

通常首先通过颜色特征发现物体,然后通过它们的形状、纹理和拓扑关系等特征来进一步识别物体。当图像中有明显物体出现时,图像的内容可以由这些物体的颜色、位置和形状等特征表示。综合颜色、形状和空间的图像检索的过程分为三步:首先对图像进行分割,得到主要物体所占的区域;然后对每一块区域提取各自的颜色、位置和形状等特征作为检索对象的特征;最后根据图像中各对象的特征计算来确定两幅图像间内容的相似程度。

在实际应用中,综合利用颜色、纹理、形状和空间关系等不同特征进行检索有许多优点。首先,可以达到不同特征的优势互补的效果。在颜色特征的基础上加上形状特征不仅能描述图像的整体颜色性质,还可以描述目标图像局部的颜色性质,而在颜色特征的基础上加上空间关系特征能较好地表达景物的结构而且相当直观。其次,可以提高检索的灵活性和系统的性能以满足某些实际应用场合的需要。综合相似性采用下式计算:

$$S=\frac{\omega_c S_c+\omega_r S_r+\omega_s S_s}{\omega_c+\omega_r+\omega_s} \tag{7-36}$$

其中，ω_c、ω_r、ω_s 分别为颜色、形状和空间的加权，S_c、S_r、S_s 分别为颜色、形状和空间的相似度。

7.5 基于视觉特征的图像检索系统

基于视觉特征的图像检索能够自动提取每幅图像的视觉特征作为其索引，如色彩、纹理和形状等，查询将根据图像视觉特征进行相似性计算。用户通过选择具有代表性的一幅或多幅用例图像来构造查询，然后由系统查找与用例图像在视觉内容上比较相似的图像，按相似性大小排序返回给用户。另外，基于视觉特征的图像检索系统一般还可以通过可视化界面和用户进行实时交互，便于用户构造查询、评估和改进检索结果。

7.5.1 基于视觉特征的图像检索系统整体架构

图 7-11 是基于视觉特征的图像检索系统的总体架构。系统的主要模块包括图像分割模块、特征选择抽取模块、索引模块、特征向量索引库、用户界面、图像检索模块、相似性度量模块、相关反馈模块和显示模块。

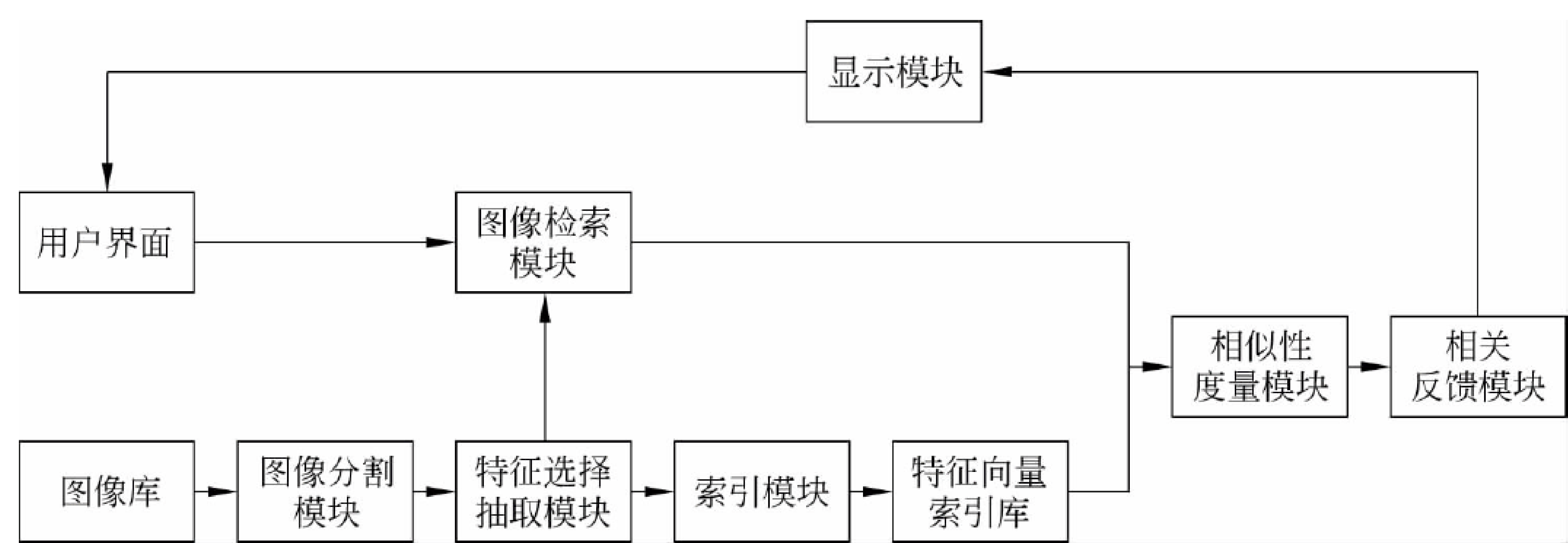


图 7-11 基于视觉特征的图像检索系统的总体架构

(1) 图像分割模块。基于视觉特征的图像检索系统可以直接抽取整幅图像的视觉特征进行相似性计算，也可以先将图像分割成一些有意义的区域后再抽取区域的视觉特征进行相似性计算。

(2) 特征选择抽取模块。特征选择抽取模块主要根据选择的特征，进行相关的图像颜色、纹理和形状等视觉特征的自动抽取，为相似性度量做准备。

(3) 索引模块。为了减少检索图像的响应时间，索引模块对抽取的图像特征向量建

立索引。索引模块不仅可以自动建立索引,而且可以对索引进行动态管理。

(4) 特征向量索引库。特征向量索引库是存储和管理特征索引的模块,可以采用关系数据库实现特征向量索引库的管理。

(5) 用户界面。用户界面的主要功能是为用户提供功能强大的搜索表达机制和灵活的搜索方式。

(6) 图像检索模块。图像检索模块的主要功能是根据用户选择的查询实例,调用特征抽取模块,抽取实例图像的特征向量,供相似性度量模块使用。

(7) 相似性度量模块。相似性度量模块将查询实例的特征向量与索引库中的图像特征向量进行相似性计算,并根据相似性的大小排序。该模块还根据相关反馈信息,重新调整参数来计算相似性,以获得更加符合用户需求的查询图像。

(8) 相关反馈模块。相关反馈模块提供人机交互的接口,模块将用户对查询结果的反馈信息返回给相似性度量模块。通过多次的人机交互与学习对话,提高检索的精度。

(9) 显示模块。显示模块实现查询结果的显示,本模块根据相似性度量和相关反馈的结果,找到原始图像,采用依据相似性排序和缩略图的方式,以图像列表或图像反馈网页的形式将结果展现给用户。

7.5.2　图像分割技术

图像分割是把图像中互不相交,具有特殊含义的区域区分出来。每个区域内的像素属性满足一定的一致性,如灰度值相近或纹理特征相似等。图像分割是图像理解的关键步骤,尽管已经有了许多分割方法,但是到目前为止还不存在一种通用的方法,同时也没有一个判断分割质量的标准,因为分割与人的主观认识有密切联系,被认为是计算机视觉图像处理中的一个瓶颈技术。

1. 图像分割的概念

图像分割是指把图像分解成各具特性的区域并提取出感兴趣目标的技术和过程。图像分割一般定义为:设 I 为一幅图像,H 是一个衡量像素属性一致性的函数,它的取值为两个:true 或 false,那么图像分割就是把图像 I 分成 n 个区域 $R_i=(i=1,2,\cdots,n)$,满足:

(1) $\bigcup_{i=1}^{n} R_i = I$。

(2) 如果 $i \neq j$,那么有 $R_i \bigcap R_j = \varnothing$。

(3) $H(R_i)=\text{true},\ i=1,2,\cdots,n$。

（4）对所有相邻的区域 R_i 与 R_j，$H(R_i \bigcup R_j)=\text{false}$。

由分割的定义可知：条件（1）和（2）指出分割是把整幅图像分成一些互不相交的区域，并且这些区域的并集是整幅图像；条件（3）指出每个区域内部的像素满足一定的属性一致性，条件（3）指出根据给定的属性一致性判断函数 H，任何两个相邻的区域不可能合并成一个区域。

2. 图像分割算法

对图像分割算法的研究已经开展了几十年，至今借助于各种理论已经提出了许多分割算法，而且这方面的研究仍然在积极推进。目前已经提出的分割算法大都针对具体的图像问题，并没有一种适合于所有图像的通用分割算法。实际上由于不同领域的图像千差万别，也不太可能存在万能的通用算法。图像的分割算法非常多，大体上可以分为以下几类。

1）基于空间特征的分割方法

这类方法的思想是：由于分割过程中考察的图像像素总具有一定的特征，因而可以把这些像素映射为一定特征空间中的点，从而将图像分割转化为特征空间中点的分类问题。常用的分类手段包括阈值化分割方法和特征空间聚类方法。

（1）阈值化分割方法。阈值化分割方法已经有几十年的历史，是图像分割领域中较早出现的一类方法，也是最基本的方法，在灰度图像的分析和识别中起着重要的作用。其目的是按照图像的灰度级，将图像空间划分成与现实景物相对应的一些有意义的区域。各个区域内部灰度级是均匀的，而相邻区域之间的灰度级是不同的，其间存在着边界。

阈值化分割技术有单阈值分割和多阈值分割。单阈值分割就是设定一个灰度阈值 T，对于一幅灰度图像 $f(x,y)$，将图像中的像素分成两类：满足 $f(x,y)>T$ 和 $f(x,y)\leqslant T$，一类称为目标，另一类称为背景。这种分割技术在机器视觉、文字识别、生物医学图像分析、指纹与印章鉴定、光学条纹判读以及军事目标识别等领域应用较为普遍。更一般地，多阈值分割则选择多个阈值，把整个灰度范围划分成几个段，隶属于每个段内的像素成为一类，这样就将图像分割成多个灰度不同的区域。显然，单阈值分割是多阈值分割的一种特殊情形。

阈值化分割技术分为两个步骤：首先是确定合适的或者是最佳的阈值，然后将图像像素的灰度和阈值进行比较，进而确定每个像素所属的类。显然，合适阈值的确定是难点和关键。阈值化分割技术中的各种各样的算法大多围绕着阈值如何选取来展开。阈值化分割技术中主要的算法包括：直方图方法和直方图变换法、最大类间方差法、最小误差法

与均匀化误差法、最大熵方法、模糊集方法、局部阈值分割与动态阈值分割及其二维阈值化方法。

阈值化分割技术比较简单直观,但它对噪声影响敏感。例如,在噪声比较严重的时候,直方图中甚至会出现虚假波峰或波谷,导致最终的结果出现明显误差。为了克服这一问题,在对图像进行直方图阈值分析前,往往需要采取适当的去噪措施,这又带来了额外的工作。另外,由于直方图并不包含空间信息,所以这类方法往往对图像空域相关性和连续性缺乏考虑,以至于分割结果的空间紧凑性一般较差。

(2) 特征空间聚类方法。特征空间聚类技术不需要训练样本,是一种无监督的全局分类方法。其中,k-均值聚类算法最为经典,它不仅应用于图像分割,还广泛应用到矢量量化和数据压缩中。另一种常用的色彩空间聚类方法是 ISODAT(interactive self-organizing data analysis technique)聚类,它是在 k-均值聚类算法基础上发展起来的聚类方法。在经典 k-均值聚类的基础上,将图像局部的自适应性和空间连续性结合起来,形成了另一类非常重要的聚类方法——自适应 k-均值聚类算法。

总的来看,特征空间聚类技术也存在一些不足:①无论是 k-均值聚类算法还是所派生出的其他方法,都存在初始的 k 个中心(或均值点)的选取问题,不恰当的初始点可能使最终的聚类结果很不理想;②绝大多数的聚类算法没有很好地考虑像素的空间位置和像素特征的空域相关性、连续性,因而分割结果在空间分布上往往不够紧凑;③自适应 k-均值聚类算法在一定程度上克服了空间问题,但其计算复杂性比较高。

2) 基于图像域的分割方法

基于特征空间的分割方法对空域连续性和相关性缺乏考虑,分割结果的连通性通常不是很理想,需要后续处理措施来改善连通性。连通性作为分割必须满足的条件之一,更好的方式是在分割的过程中就予以充分考虑。由于对象表面的连续性,同一对象的像素点在空间分布上往往很相近。基于这一事实,就必须综合考虑图像区域色彩、纹理等特征的一致性和空域分布的连续性与相关性,基于图像域的分割方法就是基于上述思想提出来的。根据所采用的空间分组策略的不同,可以把这一类方法细分为分裂-合并技术、区域生长技术、基于区域边缘检测的技术。

(1) 分裂-合并技术。分裂-合并策略的分割算法一般都是以一个不具有特征一致性和空间连续性的图像(常常是原图本身)作为初始划分,反复进行分裂过程,直到分裂出的区域都满足一致性要求;然后再执行合并过程,合并那些被过度分割的区域,从而得到最终的分割结果。分裂阶段常以四叉树为数据结构;而合并阶段则常以区域邻接关系图 RAG(region adjacency graph)为数据结构。

分裂-合并策略的分割算法一般都需要根据图像的统计特性设定图像区域特征的一致性测度以确定对一个区域是应该分裂还是合并,或者停止操作。其中最常用的做法是基于色彩的统计特性,例如同质区域中的方差(variance within homogeneous regions, VWHR),算法根据 VWHR 的数值来确定合并或分裂各个区域。VWHR 会受到图像噪声的影响,为了得到正确的分割结果,就需要根据图像中的噪声水平来选 VWHR。但图像的噪声水平一般很难准确测定,所以 VWHR 常根据先验知识或噪声估计来选定,它的选择精度对算法性能的影响很大。另外还可以借助区域的边缘信息来确定是否对其进行合并或者分裂,但其分割结果同样易受噪声的干扰。

(2) 区域生长技术。区域生长技术的基本思想是:逐个扫描图像中的像素点,找出尚未归类的像素;然后以该像素为种子,找出与其邻接的并满足预定义特征和一致性准则的像素点,合并到该种子区域;反复进行这一合并过程直至所有的像素点均被唯一地归并到某个种子区域。这一过程实际上是两个聚类过程,其结果与处理过程与像素扫描顺序有很大的相关性。一般地,区域生长技术有三个关键问题需要解决:①选择一组能正确代表区域的种子像素;②确定生长过程中合并相邻像素的特征相似性(或一致性)准则;③指定停止生长过程的条件。事实上,在区域生长停止后,经常会有一些零碎的小区域存在,因此,大多数利用区域生长的分割方法都需要采用区域合并以作为后续处理措施。

(3) 基于区域边缘检测的技术。边缘检测是图像处理领域一个研究了很长时间的问题,最早的对灰度图像边缘检测的研究可以追溯到 1965 年。现在已经有了为数众多的检测算子,它们使用不同的数学工具来实现边缘检测。利用图像梯度信息的微分算子:Laplace 算子、Sobel 算子、Roberts 算子、综合正交算子、Canny 算子等;利用数学形态学腐蚀膨胀运算的形态学算子;利用小波的小波算子等。

归纳起来,基于区域边缘检测的分割方法一般复杂度较大,这是因为边缘检测并不能直接得到图像区域,往往还需要区域填充、裂缝弥合等复杂的后续处理才能得到最终的结果。而且这类技术对噪声敏感,所以一般需要在预处理过程中采取某些去噪措施。

3) 基于模糊理论的分割方法

上述的方法在进行图像像素归类时,基本上都是以一种确定性的方式进行决策,即认为一个图像像素只可能属于一个区域,而隶属其他区域的可能性为零。事实上,由于在图像表示、分析与理解的各个层次上都存在不确定性,有时候图像中的区域并不具有明确的定义,因此图像像素的分类决策也不能明确地进行。更合理的方式应该对各层上的不确定性进行处理,并将其向更高层次传递,这样可以为高层保留尽可能多的信息,从而避免由于过早的低层判定而导致高层的决策出现偏差。

利用模糊理论的分割方法正是基于上述思想而提出的。在图像分割(边缘检测)中使用模糊集。在模糊集合中,像素属于某个区域的程度用隶属度来表示。源于模糊集合的概念,产生了模糊测度和模糊积分的概念。模糊测度用于度量模糊程度,模糊积分可以理解为模糊期望。

(1) 模糊特征空间聚类。基于特征空间聚类的分割方法中,可以将图像像素映射为特征空间中的一些点,然后通过聚类来实现点的分类。k-均值聚类算法是一种确定性方法,聚类过程中进行的是一种二值(0-1)硬决策,即一个点总是要么属于某个类,要么不属于该类。事实上,由于不确定性的存在,这样的点分类方式并不合理。k-均值聚类方法与模糊数学相结合,产生了著名的模糊 k-均值聚类方法。与 k-均值聚类一样,模糊 k-均值聚类算法的聚类结果受初始条件影响较大,而且该方法计算量比较大。针对这个问题,提出用快速模糊 k-均值聚类彩色图像分割方法来减少计算量,明显提高了模糊 k-均值聚类的计算速度。另外,把模糊积分看成是某个目标属于一个特定类的最大置信度,并将模糊积分作为山峰聚类中"距离"的测度,用于度量彩色图像数据间的相似程度。也有将模糊理论引入 Gibbs 随机场,提出了广义模糊 Gibbs 随机场,然后基于该描述模型通过聚类来实现分割,在医学图像的分割上取得了较好的效果。

(2) 模糊区域生长。模糊区域生长把区域看成是"颜色基本相同,并存在缓变化的像素集合",在 RGB 颜色空间中根据颜色向量间的欧氏距离定义了两个(相邻)像素之间对比度的隶属度函数,借以作为区域生长的相似性指标。由于 RGB 空间的三个颜色分量是彼此相关的,所以在该空间中使用欧氏距离来度量颜色差异并不合适。一种基于模糊连接度的分割方法,需要人工参与,即由用户来选定种子点,然后算法自动计算各点到种子点的模糊连接度和最优路径,最后用户通过选取阈值来得到分割结果。也有学者提出一种基于模糊颜色相似测度的彩色图像分割方法,首先在 HLS 颜色空间上定义了一个模糊颜色集,并把图像中的每一个像素都表示为一个模糊颜色集,然后利用两个模糊集合的相似测度来度量像素的相似程度,最后以该相似测度为准则反复合并相邻像素,以形成有意义的区域。

(3) 模糊边缘检测。模糊推理可方便地用于边缘检测,利用模糊推理规则产生了一种 HIS 空间中的边缘检测方法,即先利用线性的模糊隶属函数来描述两个像素在各分量上的绝对差异,然后定义若干个 3×3 的边缘结构,并使每个结构对应一条模糊规则,再根据这些规则通过推理来得出代表某个分量潜在边缘的模糊集合。推理时,一个像素可能在 1 个、2 个甚至 3 个分量上被检测出是边缘点。对每一种情况的推理结果进行加权求和,则可求得表示颜色边缘点的模糊集合。

另一种方法是利用 H 和 I 两个颜色分量进行边缘检测，分量 I 可以检测出大部分边缘，但对于相同亮度不同色调的区域，则需用分量 H。隶属度函数可通过直方图确定，同样地，六个 3×3 的潜在边缘结构对应六条模糊规则，另外，还可将边缘检测的结果和区域抽取方法相结合，以提高分割质量。

除上面提到的基于模糊理论的分割方法外，也有将模糊理论与直方图阈值技术相结合来实现分割的方法。总的来看，通过在分割过程中引入模糊理论，可以对各个层次上的不确定性进行处理，并将其尽可能保留到高层，从而不影响高层的决策。但模糊方法的引入增加了计算量，有时候甚至使算法变得十分复杂，计算开销难以接受。

4)基于特定理论工具的分割方法

除了上述几类分割方法外，图像分割领域还有一些基于特定理论工具的算法，主要有基于数学形态学的、利用神经网络的、基于小波分析和变换的、基于遗传算法的分割方法等。这些特定理论工具的算法绝大部分都是针对某个方面的具体应用提出的。

3. 分割方法存在的不足

从 20 世纪 60 年代展开对边缘提取方法的研究至今，图像分割技术已经经历了五十多年的发展，这期间研究人员提出了许多的分割方法。但是，这项技术并不成熟，还存在诸多的问题。归纳起来，当前分割技术主要存在如下一些不足。

(1) 现有分割方法一般只考虑了图像视觉特征的一致性，因而分割得到的结果通常也只是一些视觉特征一致的图像区域，与对象分割的目标相去甚远，而引入高层特征的对象分割方法尚处于起步探索阶段。

(2) 现有分割方法的准确性与通用性一般较差，分割精度亟待提高，不同的方法往往只对特定的图像和特定的应用背景有效，例如把医学图像分割方法用来分割自然景物图像一般会效果很差。

(3) 多数已有的分割方法复杂度较高，分割所需的计算时间较长，很难满足一些实时应用的需要。

(4) 缺乏通用有效的评价指标。尽管已经提出了一些定量的分割质量评价指标，但它们都存在这样或那样的问题，并没有得到普遍认可与接受；很多时候，对分割好坏的评估仍旧依赖于人眼的主观判别。

7.5.3 相似性度量

在基于视觉特征的图像检索过程中，图像的相似性本质上就是图像视觉特征的相似性。近几十年，不同的研究人员提出了许多不同的相似性度量模型。通常相似性度量应

满足以下一些性质：①与语义相吻合；②对噪声鲁棒；③计算的有效性(能够达到实时并且在高尺度的条件下能够计算)；④对背景具有不变性；⑤局部线性(在邻近区域满足三角不等式)。常用的相似性度量模型有几何模型、相关计算模型、关联系数模型等。

1. 几何模型

几何模型将图像的特征看做是坐标空间中的点，通常用两点之间的距离表示它们的相似程度。设 d 为距离度量函数，$\boldsymbol{s}_1$、$\boldsymbol{s}_2$、$\boldsymbol{s}_3$ 为三个特征向量，则距离度量函数的定义需要满足以下的距离公理。

(1) 自相似：$d(\boldsymbol{s}_1,\boldsymbol{s}_1)=d(\boldsymbol{s}_2,\boldsymbol{s}_2)=d(\boldsymbol{s},\boldsymbol{s}_3)=0$。

(2) 对称性：$d(\boldsymbol{s}_1,\boldsymbol{s}_2)=d(\boldsymbol{s}_2,\boldsymbol{s}_1)$。

(3) 三角不等性：$d(\boldsymbol{s}_1,\boldsymbol{s}_2)+d(\boldsymbol{s}_2,\boldsymbol{s}_3)\geqslant d(\boldsymbol{s}_1,\boldsymbol{s}_3)$。

常用的距离度量函数有以下几种：

(1) Minkowsky 距离。Minkowsky 距离可以延伸为 Manhattan 距离、欧氏距离和切比雪夫距离等。

$$d(\boldsymbol{x},\boldsymbol{y}) = \left[\sum_{i=1}^{N} |x_i - y_i|^r\right]^{\frac{1}{r}} \tag{7-37}$$

当 $r=1$ 时为 Manhattan 距离：$d(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{N} |x_i - y_i|$。

当 $r=2$ 时为欧氏距离：$d(\boldsymbol{x},\boldsymbol{y}) = \left[\sum_{i=1}^{N} |x_i - y_i|^2\right]^{\frac{1}{2}}$。

欧氏距离是常见的距离度量函数，具有空间不变性的特点，但欧氏距离没有考虑到各维之间的关系，所以在图像检索中较多使用加权欧氏距离。当 $r\to\infty$ 时为切比雪夫距离：$d(\boldsymbol{x},\boldsymbol{y})=\max\limits_{1\leqslant i\leqslant N}|x_i-y_i|$。

(2) 直方图相交距离。用于以直方图为特征向量的相似性度量。

$$d(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{N}\min(x_i,y_i)\Big/\min\left(\sum_{i=1}^{N}x_i,\sum_{i=1}^{N}y_i\right) \tag{7-38}$$

(3) 直方图二次式距离。两个颜色直方图 X 和 Y 之间的二次式距离可以表示为

$$D(\boldsymbol{X},\boldsymbol{Y}) = (\boldsymbol{X}-\boldsymbol{Y})^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{X}-\boldsymbol{Y}) \tag{7-39}$$

对基于颜色直方图的图像检索来说，二次式距离比使用欧氏距离或是直方图相交距离更为有效。因为它通过引入颜色相似性矩阵 $\boldsymbol{A}$，使其能够考虑到颜色相似但不相同的图像，但该方法的运算代价较大。

(4) Mahalanobis 距离。如果特征向量的各个分量间具有相关性或者具有不同的权

重，可以采用 Mahalanobis 距离。

$$d(\boldsymbol{x},\boldsymbol{y}) = \sqrt{(\boldsymbol{x}-\boldsymbol{y})^{\mathrm{T}} \sum^{-1} (\boldsymbol{x}-\boldsymbol{y})} \tag{7-40}$$

式中，$\sum^{-1}$ 为特征向量 $\boldsymbol{x}$、$\boldsymbol{y}$ 的协方差矩阵。

2. 相关计算模型

相关计算模型是计算两个特征向量之间的相关性，相关性越大，说明越相似。常用的相关方法有内积相关、余弦相关、佩尔森（Pearson）积矩相关等。

（1）内积相关

$$R(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{N} x_i y_i \tag{7-41}$$

（2）余弦相关

$$\cos\theta = \frac{\sum_{i=1}^{N} x_i y_i}{\sqrt{\sum_{i=1}^{N} x_i^2 \sum_{i=1}^{N} y_i^2}} \tag{7-42}$$

（3）佩尔森积矩相关

$$R(\boldsymbol{x},\boldsymbol{y}) = \frac{\sum_{i=1}^{N} (x_i - \overline{x_i})(y_i - \overline{y_i})}{\sqrt{\sum_{i=1}^{N} (x_i - \overline{x_i})^2 \sum_{i=1}^{N} (y_i - \overline{y_i})^2}} \tag{7-43}$$

式中，$\overline{x_i}$，$\overline{y_i}$是图像数据库中所有第 i 个特征的均值。

3. 关联系数模型

若图像中的有些特征是二值型的，则用关联系数模型计算。例如，令二值特征向量分别为 $\boldsymbol{x}_i$，$\boldsymbol{y}_i$，则 Gower 关联系数如下式：

$$S_G = \sum_{i=1}^{N} w_i s_i \Big/ \sum_{i=1}^{n} w_i \tag{7-44}$$

式中，$s_i = \boldsymbol{x}_i \oplus \boldsymbol{y}_i$，如果 $\boldsymbol{x}_i$ 与 $\boldsymbol{y}_i$ 匹配，则 $s_i=1$，否则 $s_i=0$，w_i 为权重因子。

7.5.4 图像索引

对于大规模图像数据库来说，线性扫描已经很难满足用户的需求，因此需要利用相应的技术和数据结构来组织特征向量并管理搜索过程从而加速查询，这就是索引的基本功

能。图像数据库的索引机制与一般索引结构的一个重要区别在于它面临着维度问题带来的影响。

为了使基于视觉特征的图像检索技术能够应用于大规模的图像库,必须采用有效的多维索引技术。它存在的难题有两个方面:一是高维数,通常情况下,图像特征向量的维数量级是 10^2;二是非欧拉的相似性度量,由于欧拉度量方法可能无法有效模仿人类对视觉内容的所有感知,因此经常需要采用其他的相似性度量方法,例如直方图的交、余弦、相关性等非欧拉的相似性衡量方法。近年来研究者提出了很多解决方法,它们可以分为六类:高维索引方法、降维方法、近似最近邻方法、单一维空间映射方法、多重空间填充曲线方法和基于过滤的方法。

1. 高维索引方法

高维索引方法是近几年信息检索领域的研究热点。索引机制的关键问题是如何划分数据空间,以及如何根据划分方法将数据组织起来。根据数据空间的划分方法,通常将高维索引方法分为两类:基于空间划分的方法和基于数据划分的方法。基于空间划分的方法是对数据所在的空间进行划分,这种方法主要包括四叉树、*K-D* 树、R^+ 树和网格文件等;基于数据划分的方法是根据数据对象进行划分,这种方法主要包括 *R* 树、R^* 树、*X* 树和 SR 树等。

(1) 四叉树。四叉树是一类常见的索引结构,属于基于空间划分的索引结构。在四叉树创建时,首先将整个空间划成四个相等的子空间,然后对每个或其中几个子空间再继续划分,这样就形成了一个基于树图的空间划分。四叉树是一种金字塔式的数据结构。当图像是方形的,且像素点的个数是 2 的整数次幂时,四叉树最合适。四叉树的根节点对应于整幅图像,叶节点对应各单个像素或具有相同特性的像素组成的方阵,所有的点可分为三类:目标节点、背景节点、混合节点。同 *R* 树相比,四叉树可以用顺序存储的线性表来表示索引,内存需求量小,插入和删除操作更加简单、方便,有利于查询速度的提高。但四叉树是一种非平衡树,在建立索引之前必须预先知道空间对象所分布的范围,可调节性比较差。

(2) *K-D* 树。*K-D* 树是一种 *K* 维空间中的二叉查找树,主要用于存储点数据。在 *K-D* 树建立时,数据集合向每个坐标轴投影,选取最长投影值的中值作为切割点,将整个数据集合分割为两个,分别作为节点的存储对象构成子节点,然后对每个子集进行同样的操作,直到每个集合最小。*K-D* 树是一个非平衡树,不同数据插入顺序会产生不同结构的 *K-D* 树。在 *K-D* 树中,数据不仅出现在叶节点上,也可以分散在树的任何地方。*K-D* 树虽然对存储要求比较低,但却增加了树的深度,不利于海量数据存储,树的更新也比较

困难。

(3) R^{+}树。R^{+}树的提出是为了解决 R 树中 MBR 重叠造成的多重路径搜索问题。R^{+}树采用特殊的分裂方式,分解后的对象存储在几个节点中,使同层的节点矩形之间不再存在重叠。实验表明,与 R 树相比,R^{+}树的性能有一定提高,特别是针对点查询,可以减少超过 50%的磁盘访问次数,但需要占据较多的存储空间。

(4) 网格文件。网格文件(gird-file)是一种典型的基于哈希表的数据存取方式,它是由包含很多与数据桶相联系的单元网格目录来实现的。一般一个数据桶对应于硬盘上一个磁盘页,每个单元只对应一个数据桶,而一个数据桶可以包含着几个相邻的单元。网格文件索引方法的优点是算法实现较为简单,结合编码技术可以快速实现目标查询;缺点是数据冗余较大,缺少层次,灵活性差,无法实现多分辨率。网格文件的变种主要有 EXCET、两层网格文件和 Twin 网格文件等。EXCET 与网格文件的不同之处在于其所有的网格单元大小都是相同的,因此每次分裂都将导致目录大小成倍增长。两层网格文件的基本思想是再增加一个网格文件,形成两层网格来管理目录,其中第一层称为根目录,是第二层目录的一个大致描述,以指针指向第二层目录,而第二层目录才是真正的目录,包含了指向数据页的指针。Twin 网格文件也引入了另一个网格文件,这两个文件的关系是对等的,而且每个文件都覆盖了整个空间,数据在这两个文件中的分布是动态的。

(5) R 树。R 树是空间数据索引结构中最重要的一种层次结构,许多其他数据索引方法都是在 R 树的基础上演变出来的。R 树是一种平衡树,是一种性能比较好的索引结构。其最小外接矩形(MBR)之间允许重叠,保证了 R 树具有至少 50%的空间利用率,但这种无约束的重叠,在维数比较高时很可能会导致索引次数和存储空间的大量增加,严重影响查询效率。

(6) R^{*}树。R^{*}树的创新之处在于分裂时提出了一种"强行再插入"的概念。如果一个节点溢出,就删除一定百分比的远离中心区域的目标,再按插入方法重新插入这些目标,这种改进使得 R^{*}树的空间利用率可达 71%~76%。通过与 R 树比较,R^{*}树除了建树复杂外,其性能都超过 R 树,提高了 10%~75%,而且 R^{*}树的鲁棒性也很强,适于多种数据分布情况。IBM 公司的 QBIC 图像检索系统就是采用了这种索引方法。

(7) X 树。X 树是对 R^{*}树的一种改进。同 R^{*}树相比,X 树主要做了两方面的改进:一是分裂时进行无重叠分裂;二是节点容量增大成为超节点。这种改进使得 X 树索引结构结合了层次结构的 R 树和线性的顺序索引两者的优点,成为一种比较适合高维索引的数据结构,在较高维时的检索性能超过 R^{*}树两个数量级。

(8) SR 树。SR 树的每个节点用最小外接圆(I)和最小外接矩形(MBR)共同描述。

这种方法增加了每个节点的存储空间,同时也使得 SR 树的创建较为困难,但提高了区域之间的分离性。实验表明,同 R^* 树相比,SR 树提高了邻近查询的效率。

2. 降维方法

降维方法是通过将数据点映射到更低维的空间上以寻求数据的紧凑表示的一种技术,这种低维空间的紧凑表示将有利于对数据的进一步处理。在基于视觉特征的检索中,可以通过维数缩减处理,将图像特征向量的维数降低到一定的限度,然后应用成熟的索引机制构建相应的索引结构。常用的降维方法有:基于低维投影的降维、基于数据间相似性的降维、基于分形的降维和基于神经网络的降维等。

(1) 基于低维投影的降维。基于低维投影的降维主要包括主成分分析(principal component analysis,PCA)方法和投影寻踪(projection pursuit,PP)方法。PCA 方法是使用最为广泛的线性降维方法之一,在信号处理领域,它对应着 Karhunen-Loeve(KL)变换。概括地讲,它先将数据投影到某一个主成分上,然后寻找具有最大方差的线性特征集,进而达到降维的目的。投影寻踪的基本思想是将高维数据投影到低维子空间上,寻找能反映原始高维数据结构或特征的投影,然后通过分析和研究投影数据以达到了解原始数据的目的。

(2) 基于数据间相似性的降维。该类降维方法根据原始高维数据之间的相似性直接寻找相应的低维坐标。多维尺度(multi-dimensional scaling)、随机邻居嵌入(stochastic neighbor embedding)、等同映射(isometric mapping)、局部线性嵌入(locally Li-near embedding)以及拉普拉斯特征映射(Laplacian Ei—genmaps)等算法均属于基于数据间相似性的降维方法范畴。

(3) 基于分形的降维。如果一个数据集在所有的观察尺度下均具有自相似性,即一个数据集的部分分布有着与整体分布相似的结构,称该数据集是分形的。基于分形的降维是近年来才得到关注的一类方法。采用分形的思想,可以比较准确地估计出数据的本征维,为降维提供指导性的参考。与其他方法对本征维的估计所不同的是,基于分形的方法能得到非整数值的本征维,即通常所说的分数维。关于分数维的定义,也有多种不同的描述,其中应用较广泛的是计盒维(box-counting dimension)和相关维(correlation dimension)。

(4) 基于神经网络的降维。神经网络通常用来建模输入向量集之间的关系。在基于神经网络的降维方法中,根据算法使用的不同网络结构,又可将其分为自动编码网络(auto-encoder networks)、自组织特征映射(self-organizing mapping)和生成建模(generative modeling)等。

3. 近似最近邻方法

以往的技术集中于获得精确的查询结果,然而在多媒体应用领域,“精确”的含义具有很强的主观性。首先,样本图像本身不一定精确表达用户的意图,另外,图像本身是采用视觉特征向量来近似描述,而特征向量之间的相似性程度又依赖于具体的度量方法。因此,精确的最近邻并不一定与人类的感知相一致。近似最近邻方法的目的是在获得用户满意结果的前提下,缩小查询范围,以提高系统的响应速度。大多数的近似最近邻方法集中于 ε-最近邻(ε-NN)查询,ε 是所能容忍的最大相对误差。当原始空间固有的维数很高时,ε-NN 仍不能摆脱维数问题的困扰。有学者提出一个概率近似最近邻(PAC-NN)方法,即在已知查询点距离分布的情况下,允许以一定的概率 δ 超越误差界限 ε。由于实际的数据库在特征空间中并不会呈一致分布,一些研究者利用这种特征空间分布信息进行有效的近似最近邻查询。有学者采用了基于网格的聚类方法,首先将特征空间划分成网格,对邻近的高密度单元进行合并形成聚类,然后将每个聚类中的数据进行顺序地存储,对于相似性查询,只需读入一个或几个近邻聚类,以此可以节省大量的 I/O 操作。

4. 单一维空间映射方法

由于商用数据库管理系统都支持 B^+ 树这种有效的一维索引结构,一些研究者采取了将高维空间数据映射到一维空间进行检索的方法。Berchtold 等人提出了一个数据空间的金字塔形划分方法,其查询方式为范围查询(range query),该方法采用一个类似于剥洋葱方式对数据空间进行划分,能很好地避免维数困扰问题。金字塔技术以中心点作为顶点,将 d 维数据空间划分为 $2d$ 个金字塔。每个金字塔以平行于塔基的方式划分成多个部分。将数据点在每个金字塔划分内的高度作为对该点的近似,采用 B^+ 树对其索引。也可以进一步采用不同的数据空间划分和参考点选择方法,将特征空间映射到一维空间,并利用了 B^+ 树在范围查询的基础上,逐步增加查询半径以实现 k-NN 查询。

5. 多重空间填充曲线方法

多重空间填充曲线索引方法的基本思想是:利用空间填充曲线将高维空间的数据映射到低维空间,然后利用其他索引方法对这些低维空间的数据进行处理。Hilbert R 树就是基于这一思想提出的,它选择 Hilbert 曲线作为一种高维到低维的映射,建立在这种映射之上的 Hilbert R 树把各个数据矩形的中心映射为 Hilbert 曲线上的一个值,然后把这些值按升序排列。这样,就可以获得一棵空间利用率接近 100%的 Hilbert R 树,Hilbert R 树是一种高效的高维索引结构,但这种方法是以牺牲检索准确性为代价的。

6. 基于过滤的方法

对于一致性分布数据而言,当索引结构维数超过十维时,大多数索引结构的检索性能

甚至不如顺序扫描。基于过滤的 VA-File 对原始特征向量进行近似压缩,通过对这种压缩文件的顺序扫描来对原始特征向量进行过滤,再对原始的候选向量进行验证检查,这样就可以节省大量的 I/O 操作。VA-File 的基本思想是将高维数据进行压缩和近似存储。它将数据空间划分成 $2b$ 单元,b 表示用户指定的二进制位数,每个单元分配一个位串。位于某个单元内的向量用这个单元近似代替,VA-File 本身只是这些近似体的数组。查询时,先扫描 VA-File,选择候选向量,再访问向量文件。VA-File 采用了顺序扫描的思想。如果数据分布足够密集,对数据直接进行顺序扫描有时会比扫描索引树有更高的效率;另外,VA-File 采用了二进制表示的压缩方法,减少了索引结构的存储空间,检索效率明显提高,是目前在高维情况下唯一能优于顺序查找的一类精确索引方法。对于分布比较均匀的数据而言,其检索效果要好于顺序扫描和传统的多维检索方法,而对于具有明显聚类倾向的分布数据,其检索效果则显著下降。为了增加过滤能力,VA-File 不得不采用更多的比特数进行量化描述。也可以在每一个划分单元内进一步采用极坐标方式对位于此单元内的特征向量进行近似描述,以增加过滤能力,由于极坐标描述方式与空间维数无关,当空间维数增加时,并不需要更多的描述信息。一些研究者将这种压缩技术和索引树相结合,构造出新的索引结构。例如,将 VA-File 与 R 树结合,提出了 A 树、IQ 树等。

很多方法试图解决"维度困扰"问题,其中一些方法取得了一定进展,能够获得比顺序查找更快的检索速度。但高维索引机制还存在很多的问题需要进一步研究。这些问题主要表现在以下几个方面。

(1) 多数现有的索引机制当维数超过十维时,性能急剧下降。

(2) 对高维数据进行划分时,通常认为数据是均匀分布的,或者对数据的分布进行某些假设,但这些假设通常与数据的真实分布相差甚远。

(3) 多数索引结构不支持数据库的动态更新,或者更新代价昂贵。

(4) 多数索引结构,尤其是高维索引结构,其计算复杂度很高。

(5) 多数索引结构只能处理维数固定的数据。

(6) 通常一个新的索引机制的提出只是对某一个或一类原有机制的改进,几乎没有考虑多种不同形式的有效结合。

(7) 大部分研究工作只从提高索引性能的角度来提高基于内容检索的效率,而很少考虑从改善搜索算法的性能方面着手。

7.5.5 相关反馈技术

虽然基于视觉特征的图像检索取得了一定的成果,但图像视觉特征与高层语义之间固有的“语义鸿沟”决定了仅仅从图像视觉特征这一方面着手的检索方式无法取得满意的结果。一般认为,用户倾向于在语义层次上判断检索结果的好坏。这就是说,用户所认为的好结果必然是与用户查询在语义上高度相关的。为了解决这一瓶颈,人们提出了交互式相关反馈技术,其中心思想是:将人类理解的主观性融入图像检索过程,并且给用户以评价检索结果的机会,在用户评估的基础上再进一步改进检索过程。近年来,这一研究主题已成为基于视觉特征的图像检索研究者所关注的焦点。

相关反馈技术最初起源于文本检索,是一种用来提高检索系统精度的有监督学习方法。对一个给定的查询,系统首先根据预先确定的相似性规则检索出一系列有序图像。然后,用户对这些图像标上查询相关(正例)或查询无关(反例),系统将基于这些反馈改进查询并检索出新的一系列图像提交给用户。因此,相关反馈的关键问题是如何通过分析反馈的正例和反例调整相似性度量并改进查询的质量。

尽管文本检索中的相关反馈技术研究较少,但在基于视觉特征的图像检索中却成为活跃的研究课题。导致这种现象的主要原因是基于视觉特征的图像检索的精度较低,以至于直接应用源于文本检索的相关反馈框架也能够显著地提高精度。

在一些基于视觉特征的图像检索系统中,研究者利用查询点移动技术和轴再加权技术来实现相关反馈。查询点移动技术本质上是通过使“理想查询点”移向好的样本点并远离坏的样本点来提高其评估值。经常使用的技术是 Rocchio 提出的方法,该方法操作于相关文档 D_R 和非相关文档 D_N 集合:

$$Q = \alpha Q + \beta\left(\frac{1}{N_R}\sum_{i\in D_R} D_i\right) - \gamma\left(\frac{1}{N_N}\sum_{i\in D_N} D_i\right) \tag{7-45}$$

式中,α,β 和 γ 是适当的参数;N_R 和 N_N 分别是 D_R 和 D_N 中文档的数量。MARS 系统中实现了这一技术,实验表明这种相关反馈技术能够大大提高检索性能。在轴再加权技术中,主要是给那些范例图像更加接近的特征指派更大的权重,同时给别的特征指派小的权重。MindReader 检索系统对轴再加权技术进行了改进,他们利用加权矩阵定义椭圆距离作为图像之间的相似性度量,并优化参数使得查询图像全局分散性最小化。相关反馈也可以认为是一个分类问题。首先应用用户提供的范例图像训练一个分类器,然后分类器把数据库中的图像分成查询相关的和查询不相关的两类。

7.6　典型的图像检索系统

基于视觉特征的图像检索技术已经取得了长足的发展，迄今已有许多图像检索系统面世。下面介绍一些比较有代表性的基于视觉特征的图像检索系统。

（1）QBIC。IBM的QBIC系统是第一个商品化的图像检索系统，其系统框架与技术对后来的图像检索系统有深远的影响。QBIC系统提供对图像、视频、文本和语音多种形式的多媒体信息进行检索，它支持基于例图、用户构造的草图查询，同时也支持颜色、纹理和形状等特征的查询方式。QBIC是考虑了高维特征索引的系统之一，在它的索引子系统中，首先采用KI变换来减小维数，然后采用R^*树作为多维特征的索引结构。

（2）Photobook。Photobook是MIT的媒体实验室开发的一套交互式图像数据库浏览和查询工具。它有四种应用领域的示范：纹理识别、形状识别、人脸识别和大脑形状识别。FourEyes是Photobook的扩展版本，它突出了交互式语义查询及系统学习能力，并且还应用了相关反馈技术。

（3）VisualSeek。VisualSeek是由哥伦比亚大学开发研制的基于Web的图像/视频搜索工具，它是最早的基于区域的图像检索系统。它充分利用图像与区域之间的空间关系，从压缩域中提取视觉特征，系统所采用的视觉特征是颜色特征和基于小波变换的纹理特征。为加速检索过程，采用了基于二叉树的索引算法。例如用户查找“日落”的图像，可在草图上半部分绘制成橘红色区域，下半部分绘制成蓝绿色区域。VisualSeek系统由三部分组成：图像/视频收集器、主题分类和索引器、检索器。VisualSeek提供四十多个一级类目管理图像，用户首先通过关键词检索得到初步结果，然后根据初次反馈结果，选中满意的图像作为训练样本进行相关反馈。

（4）Netra。Netra系统是在LTCSB大学Alexandria数字图书馆项目中用于图像检索的原型系统，它是基于图像分割的检索系统。利用图像区域的颜色、纹理、形状及空间关系等信息从图像库中检索相似的区域。Netra的主要特点包括采用了Gabor滤波器的纹理特征，基于神经网络的“图像词典”的构造和基于边流法（edge flow）的图像分割。

（5）Virage系统。Virage是Virage公司开发的基于内容的图像搜索引擎。其特点是提供完善的用户开发功能，如提供用于开发用户界面的工具包；提出Primitive的概念，用于支持用户定义新的图像视觉特征；支持五种抽象数据结构便于图像特征的描述；提供用户相关反馈机制。Virage已经和多种商业数据库系统进行了集成。

（6）MARS系统。MARS是多媒体分析和检索系统的英文缩写，是伊利诺斯大学分

校开发的。它是计算机视觉、数据库管理系统和信息检索多个领域交叉的系统。MARS系统的重点并不在于找到所谓“最好”的图像特征，而在于根据实际的应用环境和用户需求在检索框架中动态地组合调整各种不同的图像特征。在图像检索中提出了相关反馈的结构，并在检索的不同层次上使用了该技术，包括查询矢量优化、自动匹配工具选择和自动特征适应。

(7) RetrievalWare。RetrievalWare 是由 Excalibur 开发的一种基于视觉特征的检索系统。它使用了颜色、形状、纹理等作为查询特征。它同时还支持将这些查询特征组合起来，并可以由用户来指定各自的权重。RetrievalWare 的技术已经部分应用到 Yahoo 的 Image Surfer 图像搜索引擎中。

(8) Blobworld。Blobworld 是 UC Berkeley Computer Vision Group 开发的基于图像分割的检索系统。该系统的一个重要特点是用户可以清楚地看到图像的表示，提交查询的同时，可以定性规定所选区域和其他区域的重要程度，以及各区域的各种特征(颜色、纹理、形状、位置)的重要程度，从而使用户能清楚地理解。

(9) Simplicity。该系统是由 Stanford 大学开发的检索系统，能对图像进行语义的分类，如纹理和非纹理图、户内和户外图等。它首先从图像的像素块中抽取小波纹理、LUV 颜色特征，然后基于 k-均值聚类方法分割图像成区域，同时将分割的结果输入到分类器中以决定图像的语义类型。Simplicity 提出了 IRM(integrated region matching，区域整合匹配)的相似性度量，这一方法通过在两幅图像的各个区域之间建立多对多的映射，以减小不精确分割的影响。

7.7 图像检索技术的发展方向

7.7.1 融合人工反馈

计算机视觉模式识别系统和图像检索系统的一个基本区别就是在后一个系统中人是必不可少的部分，需要探究人和计算机的配合，这一研究已在基于内容的图像检索系统的评估中有所表现。早期的研究主要是“全自动系统”，并寻找一种“单一的最好的特征”。但这种方法并不成功，因为计算机视觉技术还达不到这个水平。近来研究重点是一些“人机交互式图像系统”和“人工反馈检索图像系统”。

7.7.2 高层语义和低层视觉特征结合

在日常生活中，人们倾向于用高层语义。然而当前的计算机视觉技术能够从图像中

自动提取的大多数是低层特征。在受限的应用中，如人脸和指纹，结合低层特征和高层语义(面部或指纹)是可能的。然而在一般的框架中，低层特征和高层语义并没有直接的联系。为了缩减这种语义上的差异，一些脱机或是在线的处理是必要的。脱机处理可以通过用监督学习、无监督学习或是结合两者来获得。这些学习工具有神经网络、遗传算法和聚类方法。一种用户交互友好的智能查询界面可以实现在线处理，这种方法允许用户对当前检索结果的评估再反馈给计算机，在 MARS 中提到的相关反馈技术是一种有效工具。

7.7.3　面向网络图像检索

万维网的扩展是令人惊奇的。每天都有成千上万的文件被存储到网上，其中有大量的图像。为了更好地组织和检索这些几乎没有限制的海量图像信息，需要探索基于网页的图像搜索引擎。Alta Vista、Inforseek 等网页经常被访问这一事实表明基于网页的图像搜索引擎是需要的。同基于文本的图像检索相比，网页上基于内容的图像搜索引擎还需要技术上的突破。

一个主要的技术难点在于把大多数系统中用的低层视觉特征索引同更多想要的语义层联系起来。通过初步的网上实验，发现主题浏览和基于文本的匹配比基于特征的搜索更流行。部分原因是因为网上的商用图像检索系统通过用户化主题目录来组织它们的图像库。通常，不同的图像检索系统专注于不同的用户群和内容。因此，索引特征和主题分类也不同，导致各个网络图像库的互用性有所欠缺。

7.7.4　图像检索性能评价与检索服务平台

当前，一些图像检索系统基于查找正确图像时的“cost of space/time”(空间资源与时间开销)来衡量图像检索性能。尽管这些准则在一定程度上能够评估系统性能，但是远不能令人满意。图像内容的主观感知特性是造成定义一致性评估准则比较困难的一个主因。也就是说，图像感知的主观特性阻碍了客观评估标准的定义。目前，需要找到一种图像检索系统评估方法。

建立一个正常的大规模图像检索服务平台同样很重要。对于图像压缩，我们常常用 Lena 图像，它能权衡不同的纹理。对于视频压缩，MPEG 研究团队提供了健全的测试视频序列。对于基于文本的信息检索，有标准化的大规模试验台。对于图像检索试验台，MPEG-7 研究团队近来开始收集测试数据。为了使图像检索服务平台获得成功，用大规模复杂图像数据去测试其可测性(包括多维索引)；用图像内容的丰富性以测试图像多种

特征的有效性和系统的整体稳定性能。

本章小结

数字图像中包含了大量有价值的信息,为了有效地利用图像中所包含的价值内容,这就要求有一种能够快速而且准确地从海量图像中查找并获取所需图像的技术,也就是图像检索技术。图像检索通常可以分为两大类,即基于文本的图像检索和基于内容的图像检索。

为了更好地理解图像检索基础知识与基本原理,首先需要掌握有关图像的一些基本知识,包括图像色彩的要素、图像属性类型与图像格式方面的知识。图像色彩的三要素指的是色彩的亮度、色调与饱和度,在表示时用红、绿、蓝为三基色。图像的三种基本类型是位图图像、矢量图图像、印刷图。图像生成或者产生的途径与形式是多样的,所以图像格式也是多样的。

图像检索一般模型主要包括的内容是图像特征(图像的色彩、纹理和形状等特征)提取、检索匹配机制(直方图距离、欧氏距离、城区距离、信息熵等)、检索者终端、相关反馈。

图像的颜色模型主要三类:一是 *HSV* 颜色模型。*HSV* 模型即色调 *H*、饱和度 *S* 和亮度 *V*,此模型可以用三维坐标系统表示;二是 *RGB* 颜色模型,我们日常见到的最普遍的颜色模型就是 *RGB* 模型,它与人眼视觉结构密切相关,它是一个三维空间模型,三个坐标轴分别是 *R*(红)、*G*(绿)、*B*(蓝)轴,组成一个单位正立方体;三是 *YUV* 颜色模型,*YUV* 颜色模型又称 YCrCb 模型,*Y* 表示亮度信号,*U*、*V* 表示色度信号。

图像颜色特征是一种全局特征,描述了图像或图像区域所对应的景象的表面性质。在颜色特征方面,颜色直方图描述了图像颜色的统计分布特征且具有平移、尺度、旋转不变性,因此通常用颜色直方图来描述颜色特征。

纹理特征是一种不依赖于颜色或亮度的反映图像中同质现象的视觉特征。它是所有物体表面共有的内在特性,纹理特征包含了物体表面结构组织排列的重要信息以及它们与周围环境的联系。

图像内容的形状是揭示物体的本质特征之一,可以针对面积(可用像素点的个数计算)、环形性(即周长×周长/面积,周长也用像素点的个数表示)、主轴方向、偏心率、圆形率、连通性、正切角等形状特征进行匹配。通常来说,图形内容的形状特征有两种表示方法:一种是轮廓特征,一种是区域特征。前者只用到物体的外边界,而后者则关系到整个形状区域。这两类形状特征的最典型方法分别是傅里叶描述符和形状无关矩。

图像空间关系特征是指图像中分割出来的多个目标之间的相互的空间位置或相对方向关系,这些关系也可分为连接/邻接关系、交叠/重叠关系和包含/包容关系等。通常空间位置信息可以分为两类:相对空间位置信息和绝对空间位置信息。

鉴于利用图像单个特征检索的缺点,可以综合利用图像的颜色、纹理、形状和空间特征的方法,计算特征提取向量。用户可以根据需要调整各个特征之间的权重关系,以便满足不同应用情况的查询。例如,综合形状特征和空间位置关系特征可以较好地处理一些二值图像。

基于视觉特征的图像检索技术能够自动提取每幅图像的视觉特征作为其索引,如色彩、纹理和形状等,查询将根据图像视觉特征进行相似性计算。用户通过选择具有代表性的一幅或多幅例子图像来构造查询,然后由系统查找与例子图像在视觉内容上比较相似的图像,按相似性大小排序返回给用户。另外,基于视觉特征的图像检索系统一般还可以通过可视化界面和用户进行频繁的交互,便于用户构造查询、评估和改进检索结果。

基于视觉特征的图像检索系统的主要模块包括图像分割模块、特征选择抽取模块、索引模块、特征向量索引库、用户界面、图像检索模块、相似性度量模块、相关反馈模块和显示模块。

为了使基于视觉特征的图像检索技术能够应用于大规模的图像库,必须采用有效的多维索引技术,这些技术包括高维索引方法、降维方法、近似最近邻方法、单一维空间映射方法、多重空间填充曲线方法和基于过滤的方法等类型。

本章思考与练习题

1. 图像检索的含义是什么? 通常分为哪两大类?
2. 简述图像色彩三要素和三基色的内容。
3. 图像有哪三种基本类型? 举例各自含义是什么?
4. 常用图像文件格式有哪些?
5. 用图示说明图像检索的一般模型。
6. 基于文本的图像检索主要存在哪些局限?
7. 基于内容的图像检索系统主要有哪些特点?
8. 图像颜色有哪些基本模型? 各自含义如何?
9. 颜色直方图与累加直方图的各自概念含义?
10. 用哪些特征量来表示图像的纹理属性?

11. 图像形状特征中的边界链码是什么？是如何表示的？
12. 形状特征提取一般应用了哪些几何性原理？
13. 简述图像形状特征提取的一般描述原理。
14. 说明近年来图像形状特征提取的主要研究进展。
15. 图像空间关系特征的含义？图像空间关系特征提取方法有哪两类？
16. 单特征图像检索各自有何不足之处？
17. 基于多特征的图像检索技术有哪几种形式？
18. 基于视觉特征的图像检索系统由哪些主要功能模块组成？
19. 图像分割的概念含义？主要有哪些图像分割技术？
20. 通常相似性度量应满足哪些性质？
21. 图像检索有哪些多维索引方法？
22. 降维方法的含义？常用的降维方法有哪些？
23. 你熟悉或接触过哪些典型的图像检索系统？请简要说明。
24. 图像检索技术的主要发展方向有哪些方面？

第8章 音频信息检索

声音是人类获取信息和沟通交流的重要媒介。科学研究表明：人类获取的信息有83％来自视觉，11％来自听觉，而其他感官（嗅觉、味觉、触觉等）获取信息量仅占6％。在信息爆炸的今天，我们熟悉的Google和Baidu等这些以文本信息（文字符号）为主的搜索引擎，在面对大量的图、文、声、视等融合性的多媒体信息检索时显得力不从心，因此多媒体检索技术应运而生。音频信息是重要的信息类型之一，在政治、经济、文化、教育等各个领域发挥着重要作用，而且数据量日益剧增，如何高效率地从海量音频信息中查询并利用所需音频信息，已成为信息用户日益迫切的信息需求。音频检索就是通过音频特征分析，利用某种相似性测度查找用户感兴趣的音频信息内容，是多媒体信息检索的重要组成部分之一，是目前国内外信息检索领域普遍关注的一个热点。

8.1 音频的特点

声音信号是通过空气或某种介质传播的连续波，用电信号表示时，在时间上和幅度上都是连续的模拟信号。需要检索的音频信息资源主要指能够被计算机处理的数字化音频（digital audio），它将在时间上和幅度上都是连续的模拟声音信号经过采样和分层处理，进行编码后得到离散数字表示的数字信号并保存下来。采样频率越高，分层数越多，数字化的信号就越能逼近原来的模拟信号。奎奈斯特采样定理指出，如果信号的带宽有限，那么只需要大于或等于带宽2倍的采样频率进行采样，所得的样本就可以恢复原始的信号。数字化音频的优点是信息传输与保存不易失真，记录的音频信息只要数字大小不改变，记录的资料内容就不会改变，并且数字化音频便于进行非线形编辑，这是模拟信号做不到的。

8.1.1 音频信息的基本特征

根据声波的特征，可把音频信息分为规则音频和不规则音频。其中规则音频又可以分为语音、音乐和音效。规则音频是一种连续变化的模拟信号，可用一条连续的曲线来表

示,称为声波。声音的三个要素是音调、音强和音色。声波或正弦波有三个重要参数:频率 ω_0、幅度 A_n 和相位 ψ_n,这也就决定了音频信号的特征。

(1) 基频与音调。频率是指信号每秒钟变化的次数。人对声音频率的感觉表现为音调的高低,在音乐中称为音高,音调正是由频率 ω 所决定的。音乐中音阶的划分是在频率的对数坐标($10\times\log^{I}$)上取等分而得的。

(2) 谐波与音色。$n\times\omega_0$ 称为 ω_0 的高次谐波分量,也称为泛音。音色是由混入基音的泛音所决定的,高次谐波越丰富,音色就越有明亮感和穿透力。不同的谐波具有不同的幅值 A_n 和相位偏移 ψ_n,由此产生各种音色效果。

(3) 幅度与音强。人耳对于声音细节的分辨只有在强度适中时才最灵敏。人的听觉响应与强度成对数关系。一般的人只能察觉出 3 分贝的音强变化,再细分则没有太多意义。我们常用音量来描述音强,以分贝(dB=20log)为单位。在处理音频信号时,绝对强度可以放大,但其相对强度更有意义,一般用动态范围定义:动态范围 $=10\times\log(I/I_0)$(dB),其中 I 为信号的最大强度,I_0 为信号的最小强度。

(4) 音宽与频带。频带宽度或称为带宽,它是描述组成复合信号的频率范围。

音频作为一种信息载体,可以分为三种类型:一是语音,它具有字词、语法等语素,是一种高度抽象的概念交流媒体,语音通过识别可以转换为文本,文本是语音的一种脚本形式;二是音乐,具有节奏、旋律和声音等要素,是人声和乐器音响等配合所构成的一种声音,音乐可以用乐谱表示;三是波形声音,即对模拟声音数字化而得到的数字音频信号,它可以代表语音、音乐、自然界声音和合成音响。我们人耳能够听见的音频频率范围是60Hz~20kHz,其中语音频率大约分布在300~4000Hz,而音乐和其他自然声响则是全范围分布。

8.1.2 音频信息的内容层次

音频内容从整体上看可以划分为三个等级:最底层的物理样本级、中间层的声学特征级和最高层的语义级(如图 8-1 所示)。在物理样本级,音频内容是以媒体流的形式存在的,其中包含原始音频数据和数字数据(如采样频率、量化精度和压缩编码方法等)。用户通过音频录放与编辑软件如 CoolEdit 等以时间为单位(单位可以是毫秒、秒、分或时)来检索和浏览音频内容。中间层是声学特征级,声学特征是从音频数据中自动抽取的,它可以分为物理特征(physical feature)和感觉特征(perceptual feature)。前者包括音频的基频、幅度和共振峰结构等,后者表达用户对音频的感知,例如音调、响度和音色等。感觉特征一般都在某些物理特征之间存在一定的联系。最高层是语义级,它是音频内容、音频

对象的概念描述。具体来说，在这个级别上，音频的内容可以是语音识别、辨别后的结果（文本）、音乐旋律和叙事说明等。

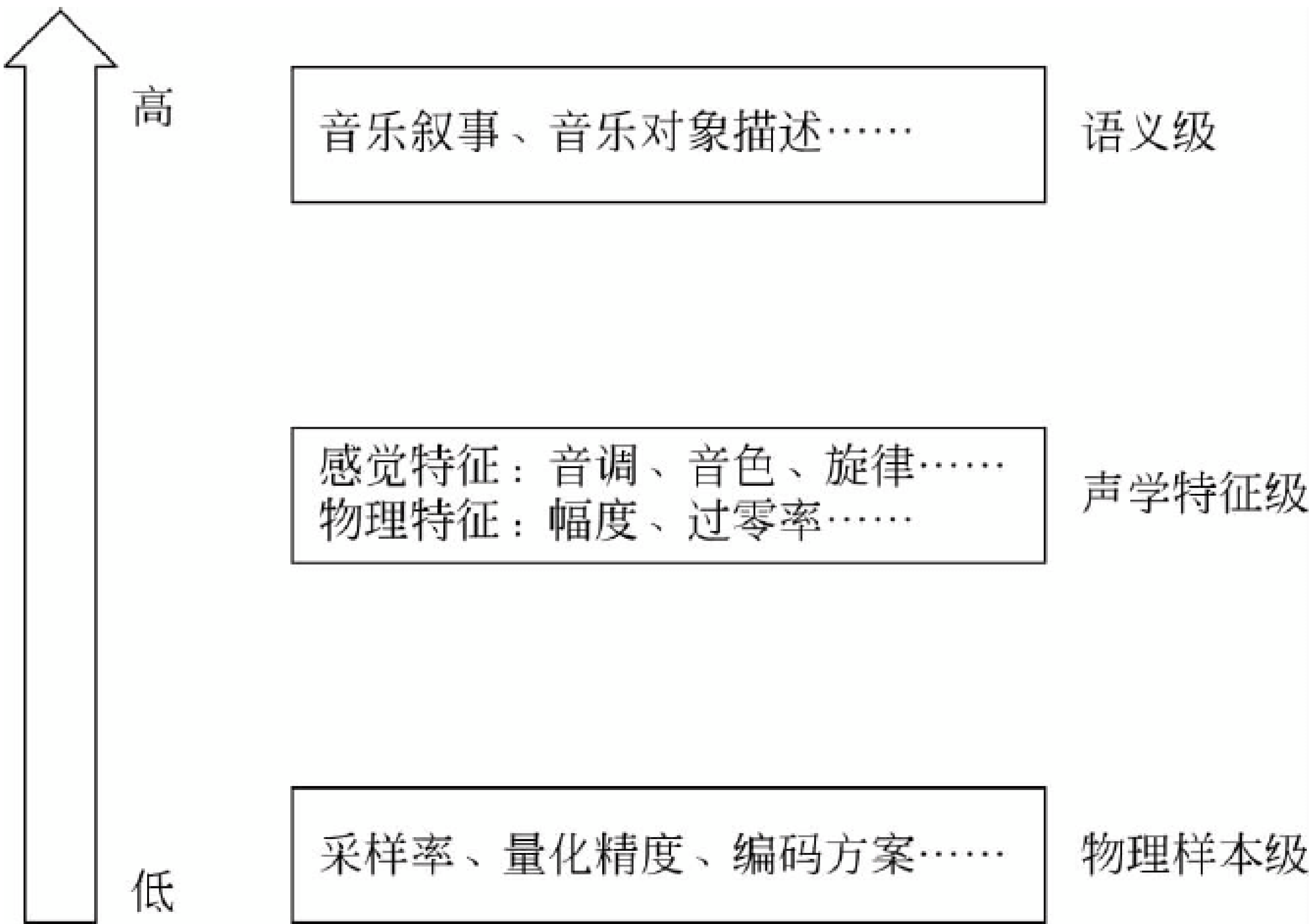


图 8-1　音频内容的抽象层次

8.2　音频信息检索技术的分类和发展

音频信息可以划分为语音、音乐和波形声音三种类型，相应的检索处理方法也可以分为以下三种：①语音检索，即以语音为中心的检索，采用语音识别等处理技术，例如电台节目、电话交谈、会议录音等；②音乐检索，即以音乐为中心的检索，利用音乐的音符和旋律等音乐特性来检索，例如检索乐器、声乐作品等；③音频检索，即以波形声音为对象的检索，这里的音频可以是汽车发动机、雨声、鸟叫等各种声音，也可以是语音和音乐等，这些声音都统一用声学特征来检索。

8.2.1　基于文本的音频检索

基于文本的音频信息检索是利用若干关键字（例如音频类型、音频标题、音频含义的文本内容描述的关键词等）组成的查询来发现匹配的音频文档。而音频信息作为一种不透明的位流，虽然可以赋予名字、文件格式、采样率等外部属性，但是首先想到的一种可行的音频检索方法是通过人工输入的属性和描述，将音频转化为文字进行检索。这种方法进行语音检索时效果显著，语音是一种特殊类型的音频，可以与文本互相转换，因此可以

利用传统文本检索方法进行概念检索，获得更准确的检索结果。

基于文本的音频检索主要借鉴了传统的文本检索技术(例如文本分类与索引、概率检索等)，在实践应用方面是盛行的，也受到了大多数音频信息检索用户的喜爱。而且在获取音频信息时，普通信息用户不需要专业检索知识，与获取文本信息的检索方法和检索习惯大体一致就可以满足音频信息的检索需要。例如，以“百度音乐”为例就可以证明这一传统检索技术的优势。见图 8-2。



图 8-2　基于文本检索的“百度音乐”实例图

图 8-2 从实践上，对于通过目录导航方式去检索流行榜单音乐的作用是十分明显的。如果需要从流行音乐的分类、歌手、专题、歌名甚至歌词或音乐专辑等方式进行检索，则提供统一的文本检索接口。见图 8-3。

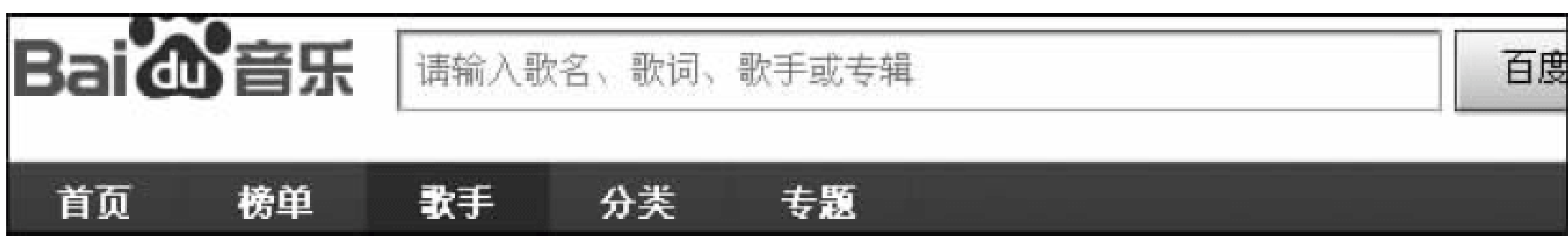


图 8-3　基于文本检索的“百度音乐”用户查询接口界面图

由于基于文本检索技术在第 2 章到第 6 章的各章中均有详细阐述，本章不再赘述。

8.2.2　基于内容特征的音频检索

基于文本的音频检索方法虽然有传统优势，但其缺点也很突出：一是当数据量越来越大时，人工注释工作量加大；二是人对音频的感知，例如音乐的旋律、音调、音质等有时难以用文字表达清楚，人工标识信息存在不完整性和主观性；三是不能支持实时音频数据流的检索。为解决上述问题，基于内容的音频检索应运而生。进行文本检索时主要提取文本的关键字等特征，进行图像检索时主要提取图像的颜色等特征，进行视频检索时主要提取视频的关键帧等特征。与此类似，基于内容的音频检索（content-based audio retrieval）就是通过从音频数据中提取和分析音频特征信息，对不同音频数据赋予不同的语义，使具有相应语义的音频在听觉上保持相似。基于内容的音频检索基本系统结构如图 8-4 所示。

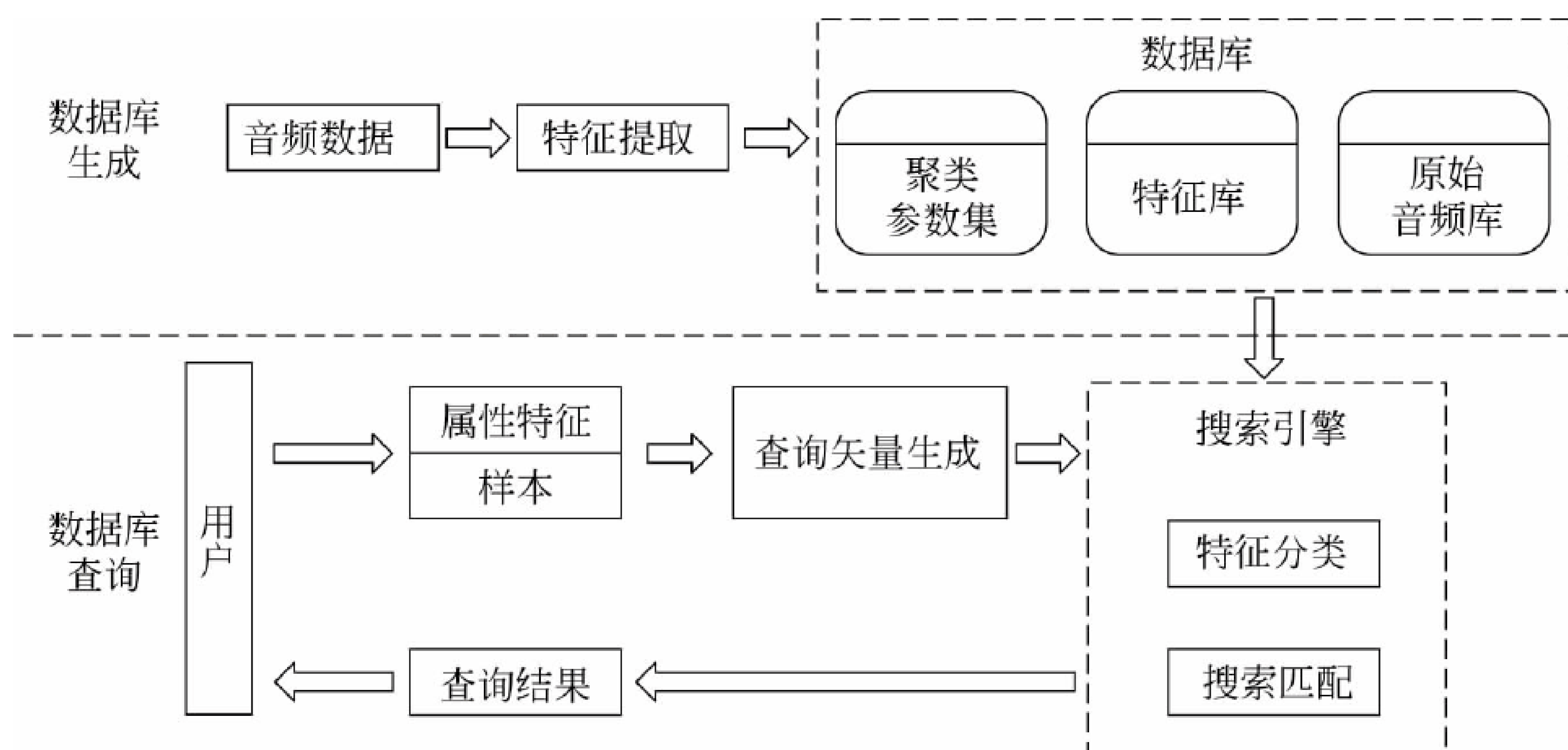


图 8-4　基于内容的音频检索系统结构

系统首先对音频数据进行特征提取，将音频数据装入原始音频库，同时将特征装入特征库。通过特征对音频数据聚类，将聚类信息装入聚类参数库部分。用户主要采用示例查询（query by example）方式进行检索，通过查询界面确定样本并设定属性值，系统接收查询后，对样本提取特征，结合属性值确定查询特征矢量，然后检索引擎对矢量与聚类参数集合进行匹配，按相关性从大到小的顺序在特征库和原始音频库中检出一定数量的相应数据，并通过查询接口返回给用户。其中原始音频库存放的是音频数据，特征库存放着音频的特征数据，按数据记录存放，聚类参数库是对音频特征进行聚类所得的参数集。

8.3 音频信息检索架构与模型

8.3.1 音频信息检索架构

基于内容的音频信息检索构架见图 8-5。

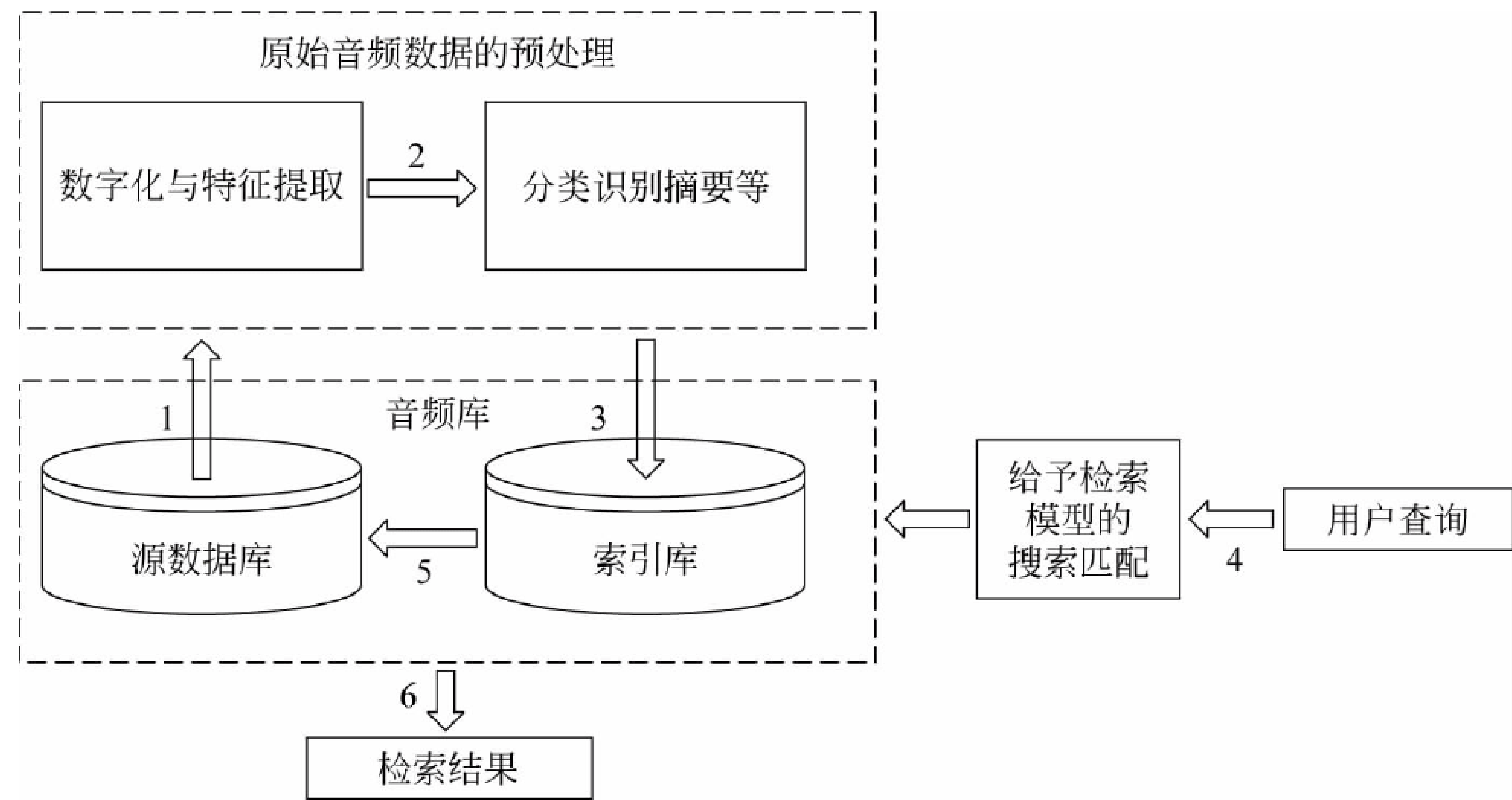


图 8-5 基于内容的音频信息检索构架

首先是索引的构建,通常是使用各种音频处理技术,先对原始音频数据进行数字化和特征提取,获取其在不同层次上的抽象信息。例如使用语音识别技术获取音频数据的语义,使用音频分类技术得到不同音频数据的类别信息,结合自然语言处理方法归纳总结音频篇章或段落的摘要等。利用上述这些信息可构造不同类型的索引库,通过索引库可快速检索到所需内容。检索时,根据用户的查询请求,通过检索模型利用索引库找到查询请求与音频库中的相似部分作为检索结果。

从上面音频信息检索的框架可以看出,音频信息检索模型在信息检索中处于非常重要的位置。所谓音频信息检索模型,就是在对音频信息进行抽象表达的基础上,通过构建一种评测机制能衡量用户查询请求与待检音频信息的相似度,即提供一种衡量用户查询请求与音频数据相似性的方法。通常可采取两者之间的距离或相似度概率来体现它们之间的相似性程度。例如,采用距离的方法,两者间的距离越近,说明它们的相似性越高,被检索出来作为结果的排序越靠前;反之,被检索出来作为结果的排列越靠后。如果用户查

询请求与待检音频没有任何相似性，则非相关音频不会出现在检索结果中。

在各种类型的多媒体信息检索中，基于文本的信息检索技术是重要基础，其检索模型也较为成熟。目前的音频信息检索技术，其模型很大程度上借鉴了文本信息检索模型的思想。典型的模型包括向量空间模型和概率模型。

8.3.2　向量空间模型借鉴

向量空间模型是一种基于统计方法的数学模型，它将请求与待检文档都表示成向量的形式。由于都是从原点出发向某个方向延伸的射线，因此空间中的各个向量间存在着一个夹角，可以使用这个夹角来度量两个向量间的相似度。一般使用这个夹角的余弦值来计算向量间的关系，两射线夹角越小相似度越高。在向量空间中，查询请求也以一条射线来表示，这条射线离哪个文档的向量射线越近，则其夹角越小，说明与查询请求越相关，检中的可能性就越大；反之就越不相关。查询请求与文件集合中的所有文档都可以计算出一个相似度，然而不能将所有文档集合中的内容都以检索结果的方式呈现给用户，因此需要设定一个阈值，根据待检文档中的内容与查询请求的相关度排序，只将排序后相关度大于阈值的内容作为检索结果。

在向量空间模型中，文档的内容被简单看成是它所含有的基本语义单位所组成的集合。将这些基本的单位统称为特征项，而原始数据文档就可以用特征项的集合来表示，记为 $D(T_1,T_2,\cdots,T_s)$，其中 T_m 是特征项，$1\leqslant m\leqslant s$。对于特征项的集合中的每一个 T_m，其在文档中的重要程度并不相同，可以赋予 T_m 一定的权重 W_m 来表示其重要程度的大小。此时文件 $D=D(T_1,W_1;T_2,W_2;\cdots;T_s,W_s)$，简记为 $D=D(W_1,W_2,\cdots,W_s)$。如果忽略特征项 T_m 在文档中的先后顺序，并要求 T_m 无异（即没有重复），就可以把 $T_1,T_2,\cdots,T_s$ 看做是一个 s 维坐标系，而 $W_1,W_2,\cdots,W_s$ 为在这个坐标系中表示文档内容的坐标值，即 $D(W_1,W_2,\cdots,W_s)$ 为 s 维空间的一个向量，称其为文档 D 的向量表示。这样查询请求与文档之间相关度评价，就可以借助查询的向量表示 Q 和文档的向量 D 来计算。

采用向量空间模型来设计有效的检索方法，需要解决以下三个问题：①如何选择特征项；②如何计算特征项的权重；③如何计算查询向量与文档向量间的相似度。

至于特征项的选择，一般用那些能够完整表示一个语义范畴的单位来作为特征项，因为这样的特征项对文档内容有较高的表示能力。如在文本检索中，常采用词和短语等作为特征项。

在文本检索中，最常采用的特征项权重计算方法是 TF-IDF 方法，其基本思想是根据特征项在文档中出现的次数来度量权重的大小，计算时要用文档长度来规定。其中 TF

(term frequency)是特征项频率，IDF(inverse document frequency)是反比文档频率，一般特征项 T_m 的反比文档频率计算如下：

$$\mathrm{IDF}_m = \lg(F/s_m) \tag{8-1}$$

其中，F 为文档集中文档的总数目；s_m 为其中含有 T_m 的文档数目。IDF 反映了这样的一个思想：如果在大多数文档中都出现的特征项的区分能力弱，所以应给以较低的权重；反之，在少数文档中出现的特征项区分能力较强，应给以较高的权重。

TF-IDF 的计算方法综合了 TF 和 IDF，采用二者的乘积作为特征项权重。文档 D_i 中的特征项 T_m 的 TF-IDF 权重计算如下：

$$W_{im} = \mathrm{TF}_{im} \cdot \mathrm{IDF}_m \tag{8-2}$$

其中，TF_{im} 为文档 $\boldsymbol{D}_i$ 中 T_m 的特征项频率。

查询向量 $\boldsymbol{Q}$ 和文档向量 $\boldsymbol{D}_i$ 之间的相似度 $\mathrm{SIM}(\boldsymbol{Q},\boldsymbol{D}_i)$ 可以采用向量内积来计算如下：

$$\mathrm{SIM}(\boldsymbol{Q},\boldsymbol{D}_i) = \sum_{m=1}^{s} W_{qm} W_{im} \tag{8-3}$$

其中，W_{qm} 为查询向量中 T_m 的权重。或者用夹角的余弦来表示如下：

$$\mathrm{SIM}(\boldsymbol{Q},\boldsymbol{D}_i) = \cos\theta = \frac{\sum_{m=1}^{s} W_{qm} W_{im}}{\sqrt{\left(\sum_{m=1}^{s} W_{qm}^2\right)\left(\sum_{m=1}^{s} W_{im}^2\right)}} \tag{8-4}$$

8.3.3 概率模型借鉴

概率模型是一种基于概率论原理，用于解决相对不确定性的信息检索模型。经典的基于概率的信息检索模型，主要依据查询请求与文档的相关度是高于还是低于非相关度的概率来进行检索。其基本思想是：给定一个用户查询，检索系统中存在着一个与该查询相关的理想命中结果集合(用 S 来表示)。如果能已知集合 S 的主要特征及描述，则用户的检索要求便不难实现。然而，在用户提供检索要求时，并不知道这个理想结果集合的特性。为此，需要在检索开始时针对 S 的特征性进行某种猜测。根据初始的猜测，系统将检索到一个初步的命中结果集合。在此基础上，用户可以对初始检索结果中文档相关与否进行判断，或者系统对检索结果文档的相关性进行自动判断。根据这些反馈信息，系统便可以在后续的检索处理中不断做出优化与改进，从而在此交互操作后使检索结果逐步接近该查询的理想命中结果 S。

如果某个文档存在几个子段，则将计算得到的最大概率值作为查询请求与被检索文档之间的相似度值。每个被检索文档按得到的相似概率值递减排列形成检索结果。

对概率信息检索模型，在实际使用中也可以只根据查询请求与文档的相关度，通过与设定的阈值进行比较来确定检索结果。进一步地，基于贝叶斯定理可以使用先验概率代替概率来进行相似度计算。

假设查询请求为 Q，第 i 个被检索文档为 D_i，则通过计算后验概率值 $P(D_i|Q)$来判断两者之间的相似度。根据贝叶斯公式

$$P(D_i \mid Q) = \frac{P(Q \mid D_i)P(D_i)}{P(Q)} \tag{8-5}$$

由于对不同的文件 $P(Q)$是固定的，因此两者之间的相似性判断也可以采用如下的公式来计算：

$$P(D_i \mid Q) = P(Q \mid D_i)P(D_i) \tag{8-6}$$

又由于对每个文档通常可以认为其出现的概率 $P(D_i|Q)$值均相同，因此在进行相似性比较时也可以忽略其影响，这样进一步近似为

$$P(D_i \mid Q) \approx P(Q \mid D_i) \tag{8-7}$$

也就是用先验概率 $P(Q|D_i)$来获得后验概率 $P(D_i|Q)$。因此，先验概率计算在概率模型中起到了重要作用，一旦先验概率得到，就可以使用它度量查询请求与被检索文档的相似性。而先验概率的计算与概率模型的选择有关，一旦模型确定，就可以使用训练得到的模型参数去计算先验概率。

8.4　表示级的音频检索

8.4.1　基于直接匹配的音频样例检索

1. 基于分段的实时音频检索

音频样例检索既可以应用于检索静态音频数据库，也可以应用于检索实时音频流。相对而言，检索实时音频流难度更大、要求更高，算法需要更多地考虑资源开销和计算速度问题。实时音频流有其自身的特点：实时性强、流过的数据无法重现，且事先不能预知，如实时广播中的电视信号数据。因此，检索必须实时地获取音频数据、计算特征、更新检索模型，然后进行匹配计算。由于实时音频流具有不可预知性，因此无法利用索引方法实现快速检索。

在音频样例检索中，通常将检索目标的音频数据作为一个整体直接检索。在整体直

接检索方法中,当输入数据流中的样例模板发生部分缺失时会增加检出的难度,甚至无法检出。而在实时检测中,流过的数据又无法再现,一旦检索时错过,无法像静态音频检索那样重现检索。在基于向量序列匹配的整体直接检索算法中,计算代价往往随着样例模板长度的增加呈线性增长。因此,当样例模板较长时,整体直接检索方法不能满足实时与快速的应用要求。同时,在实时音频检索中,必须实时计算音频特征以便更新检索模型,因此不能使用计算复杂度大的数据模型。

2. 基于 MPEG-1 压缩域模糊分类的音频检索方法

采用一种基于距离的模糊分类法,用隶属度刻画音频片段与类别之间的联系,认为每个音频片段与各个类别中心都有一个隶属关系,对不同类别之间有交叉的数据进行有效分类,解决"亦此亦彼"的现实问题,使分类结果更符合心理声学的听觉特征。

在音频分类中,人工将音频分为静音、纯语音、纯器乐音乐、歌曲、清唱、纯噪声、有背景的语音、有噪声的语音、有噪声背景的音乐等类别,设类别数为 n_c。然后用统计的方法得到各类别中心的窗特征向量,记为$\{FC_i \mid i=1,2,\cdots,n_c\}$。

若第 k 窗音频数据的特征向量为 FW_k,则它对各类的隶属度值为

$$\mu_k(j)=\frac{\dfrac{1}{\| FW_k - FC_j \|^{2/(b-2)}}}{\sum_{i=1}^{n_c}\left(\dfrac{1}{\| FW_k - FC_i \|^{2/(b-2)}}\right)},\quad (1\leqslant j\leqslant n_c) \tag{8-8}$$

其中,参数 b 用来决定对距离加权的程度。从而,可得到第 k 窗音频数据的隶属度向量 $\mu_k(\mu_k(1),\cdots,\mu_k(n_c))$。

基于隶属度的模糊分类结果明显优于硬分类方法,它可推广到 MPEG-1 所有单个层次的编码方案。检索时,用户首先在客户端提交一个 MPEG-1 音频文件或文件中节选的一小段音频片段作为要查询的音频样例,记为 $Q=\{q_i \mid i=1,2,\cdots,n_{\text{Query}}\}$,其中 Q 的长度 n_{Query}为 Q 中包含的组数量。将查询音频样例划分成 n_{win}个窗,表示为

$$n_{\text{win}}=\left[\frac{n_{\text{Query}}}{m}\right]+1 \tag{8-9}$$

其中,$\left[\dfrac{n_{\text{Query}}}{m}\right]$为向下取整算子。计算每一个窗的特征,便可得到音频样例的窗特征序列 $FW_Q=\{FW_i \mid i=1,2,\cdots,n_{\text{win}}\}$。

利用以上两个公式计算音频样例中每一个窗对各音频类别的隶属度,得到隶属度向量序列 $\mu_Q=\{\mu_i^Q \mid i=1,2,\cdots,n_{\text{win}}\}$。

相应地,在接收输入的实时流媒体音频数据时,从 MPEG-1 音频帧中提取每一组的

比例因子。当接收到与样例长度相等的 n_{Query} 组输入流数据后，同样将其划分成为 n_{win} 个窗，并计算各窗的特征向量进行模糊分类，从而得到输入音频流的隶属度向量序列 $\mu_S=\{\mu_i^S\mid i=1,2,\cdots,n_{\text{win}}\}$。输入音频流与查询音频样例的类别相似度定于如下：

$$\text{Sim}(\mu_Q,\mu_S)=\frac{1}{n_{\text{win}}}\sum_{i=1}^{n_{\text{win}}}\sum_{j=1}^{n_c}\min\{\mu_i^Q(j),\mu_i^S(j)\} \tag{8-10}$$

基于 MPEG-1 压缩域模糊分类的流媒体音频检索方案如图 8-6 所示，可以从流媒体数据中快速检索到多个任意长度的音频信息。

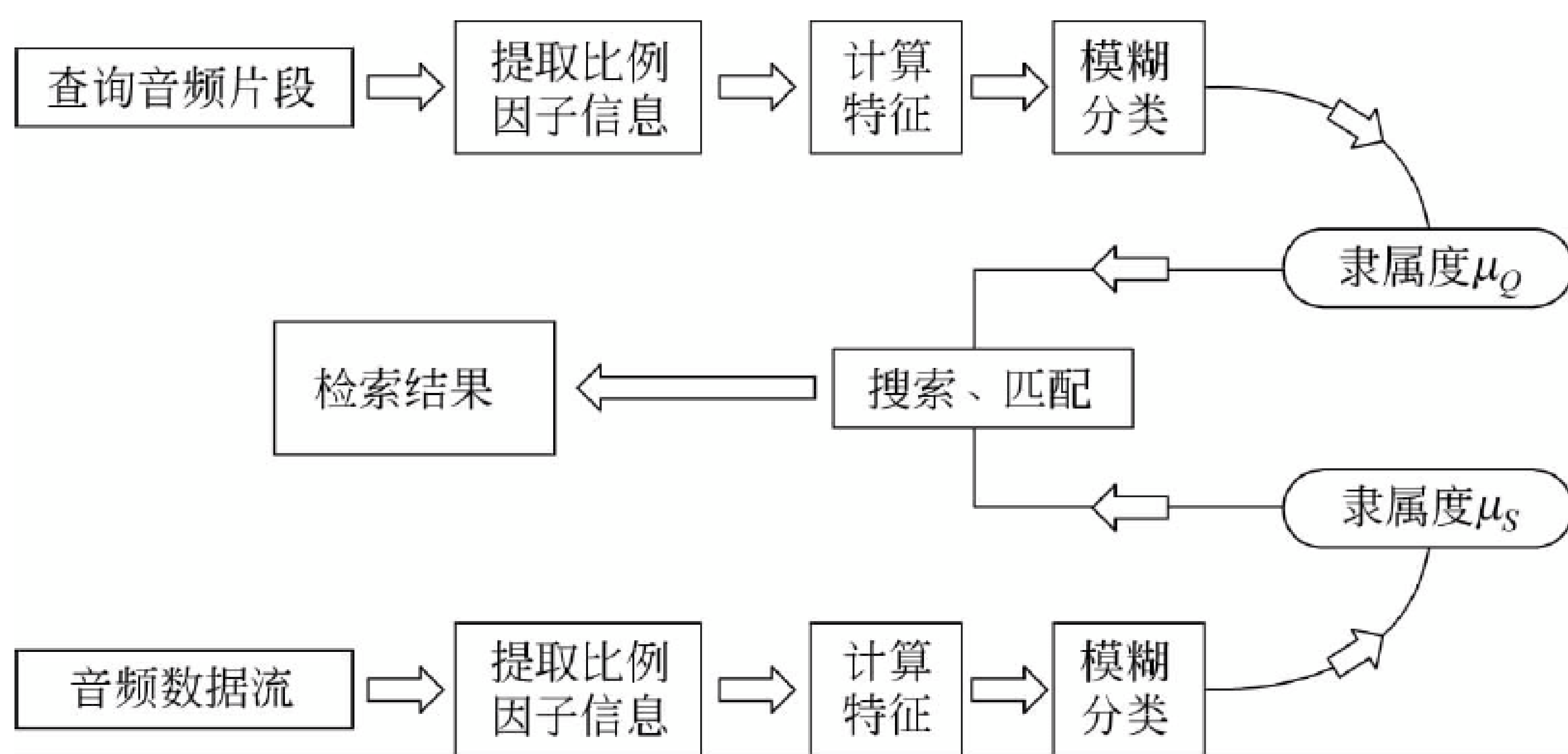


图 8-6　基于 MPEG-1 压缩域模糊分类的流媒体音频检索方法

8.4.2　基于索引的音频样例检索

从高维空间的角度来看，检索过程就是给定任意一个查询点(向量)，在数据库中找到与查询接近的点，并能保证以较高的概率返回与查询最接近的点。从概念上讲，这就很容易通过穷举法来实现：计算数据库中的所有点与查询之间的距离即可选出接近的点。如果数据库的规模很大或数据的维数很高，穷举法的实践代价往往无法接受。因此，需要寻找不依赖于数据库线性搜索的检索方法。

高维数据库的索引存在“维数的魔咒”问题，即索引的复杂度随维数的增加呈指数增长，这一直是索引研究领域中的难点。音频数据经过分帧和特征提取后得到的特征数据不仅维数高，而且还有时序信息。这就要求音频索引不仅要解决数据维数高的问题，还要在索引中体现时序性，这就进一步增加了音频数据索引构建的难度。

1. 局部敏感哈希索引方法

为解决维数魔咒问题，许多学者提出了各种检索构建算法。但无论在理论或者实践

上，这些算法与顺序扫描相比效率提升很小，在一定情况下效率甚至低于顺序扫描。近年来，为解决高维向量搜索问题，人们开始关注近似搜索问题，这是因为在大多数情况下，近似最近邻搜索具有与确切搜索同样好的结果，尤其是当距离度量准确捕捉用户的需求时，两者之间细微的差别就显得不那么重要了。

LSH(局部敏感哈希)方法产生于20世纪末，近年来这种方法开始应用于音频信息检索。LSH的思路是：如果空间中的两个点距离很近，经过投影后，这两个点的投影也会比较接近。

设 d 是 n 维空间 S 的测度，如果从空间 S 到 U 的映射族 $H=\{h:S\rightarrow U\}$ 对于任意的 $v,q\in S$ 满足以下两个条件，则该映射族是 (r_1,r_2,p_1,p_2) 敏感的。

(1) 如果 $v\in B(q,r_1)$，则概率 $P[h(q)=h(v)]\geqslant p_1$。

(2) 如果 $v\notin B(q,r_2)$，则概率 $P[h(q)=h(v)]\leqslant p_2$。其中，当 d 是距离测度，$B(q,r_1)=\{v:d(q,v)\leqslant r\}$ 时，$p_1>p_2$ 且 $r_1<r_2$；当 d 是相似测度 $B(q,r_1)=\{v:d(q,v)\geqslant r\}$ 时，$p_1>p_2$，且 $r_1>r_2$。

2. 基于局部敏感哈希倒排索引的检索方法

倒排索引(inverted index)，也称为倒排文件(inverted file)，是大型信息检索中使用最广泛的文件索引方法。所谓“倒排”表示依据检索属性来列举相关文件，是基本的信息查询方法之一。由于其快速高效的特性，倒排索引在当今谷歌、百度等成熟的基于文本搜索引擎中被广泛使用。然而，在基于内容的音频检索中，由于音频特征具有高维非字符的特性，很难直接将倒排索引技术应用其中。但如果在LSH对音频片段向量量化结果的基础上构造倒排索引，则有望获得更好性能。

(1) 倒排索引简介。所谓倒排索引，是描述一个词项集合(TERMS)元素和一个文档集合(DOCS)元素对应关系的数据结构，记为

$$\text{DOCS}=\{d_1,d_2,\cdots,d_D\} \tag{8-11}$$

$$\text{TERMS}=\{t_1,t_2,\cdots,t_D\} \tag{8-12}$$

其中，D 为文档集合大小；M 为词项集合的大小。当以“文档”为出发点时，可以称文档 d_i 中包括某些项 t_j，或者词项 t_j 在文档 d_i 中出现了多少次。而“倒排索引”直接给出的是一个 t_j 出现在哪些 d_i 中，进而还可以有它在 d_i 中出现在哪些位置。用 $\text{PL}(t_j)$ 表示 t_j 出现于其中的文档记录的集合，称为对应于 t_j 的倒排表(inverted list)，下面是关于倒排索引的几个相关量。

① E：文档集合的大小。

② $s_j=|\text{PL}(t_j)|$：词项 t_j 在文档中出现的个数。

③ $DP(t_j)=\frac{s_j}{E}$：词项 t_j 在文档中出现的频率。

④ $IDF(t_j)=-\lg DF(t_j)$：倒置文档频率，又称反文档频率，其值越小表示出现频率越高。

⑤ $f_{i,j}$：第 j 个词项 t_j 在第 i 个文档 d_i 中出现的次数。

⑥ $T_i=\sum_{j=1}^{M} f_{i,j}$：第 i 个文档 d_i 中包含的所有词项的个数。

⑦ $TF_i(t_j)=\frac{f_{i,j}}{T_i}$：词项 t_j 在第 i 个文档 d_i 中出现的频率，即词频。

⑧ $ITF_i(t_j)=-\lg TF_i(t_j)$：倒置词频，值越小表示词项出现的频率越高。

从数据结构上看，倒排文件分为两个部分：第一部分是由不同词项组成的索引，称为词表；第二部分由每个词项出现过的文档集合构成，称为记录文件，每个词项的对应部分成为倒排表，也称为记录表，可以通过词表访问。

(2) 基于局部敏感哈希的倒排索引。在 LSH 方法中，每个片段向量经过 L 个哈希函数映射后生成 L 个哈希值，一个音频文件若有 m 个片段，则将被 LSH 映射成 L 组哈希函数值序列，亦即 L 组桶号序列。见图 8-7。

$g_1(v_1)$	$g_1(v_2)$	$g_1(v_3)$	$g_1(v_4)$
$g_2(v_1)$	$g_2(v_2)$	$g_2(v_3)$	$g_2(v_4)$
⋮	⋮	⋮	⋮
$g_L(v_1)$	$g_L(v_2)$	$g_L(v_3)$	$g_L(v_4)$

图 8-7　有 m 个片段的音频文件经 LSH 映射后的结果

将桶号设置上界和下界，分别为 $Bucket_{max}$ 和 $Bucket_{min}$，超过界限的桶号一律当做界限值处理，即

$$h'_{a,b}(x)=\begin{cases} Bucket_{max}, & h_{a,b}(x)\geqslant Bucket_{max} \\ h_{a,b}(x), & Bucket_{min}<h_{a,b}(x)<Bucket_{max} \\ Bucket_{min}, & h_{a,b}(x)<Bucket_{min} \end{cases} \tag{8-13}$$

(3) 基于局部敏感哈希倒排索引的搜索。在倒排索引中，无法体现词项与词项之间的顺序关系，即无法体现音频的时序性。为了解决这一问题，可以使用一种在查找文档时所有查询词项邻近的策略，即邻近搜索(proximity search)。k 词邻近搜索方法试图找寻

在所有文档中使k个查询词全部靠近的区域。算法的复杂性既不取决于查询词之间在文档中的最大间隔距离，也不取决于查询词的个数k。使用这种技术，可以在倒排索引检索结果的排序中考虑查询词在文档中的邻近关系，从而能在一定程度上兼顾音频的时序关系。

若用$T=T[1,\cdots,F]$表示一个词项数为F的文档，$\text{Key}_1,\cdots,\text{Key}_k$表示给定查询关键词，$p_{ij}$表示文档$T$中查询关键词$\text{Key}_i$第$j$次出现的位置，则初级的$k$词邻近搜索定义如下：当给定$k$个关键词$\text{Key}_1,\cdots,\text{Key}_k$和它们在文档$T=T[1,\cdots,F]$中的位置$p_{ij}$，邻近搜索就是要在$[1,\cdots,F]$中找到区间$[l,r]$，这个长度为$r$-$l$的区间包含了所有$k$个关键词，其中关键词的顺序在区间中是随意的。

由于不知道关键词在文档中的顺序，因此在上述方法中，区间内关键词的顺序可以是随意的。而关键词以固定顺序出现的区间，则是搜索结果集合的一个子集。如果一个区间不存在包含所有k个关键词的子区间，则称它是最小区间。当k个关键词的总数为n时，区间的数量为$n(n-1)/2$，其中最大部分区间是无用的，只需找到其中包含所有关键词的最小区间即可。在k词邻近搜索中有两种算法：一种是基于平面扫描算法(plane-sweep algorihm)，另一种是基于分治的方法(divide and conquer approach)。

① 基于平面扫描算法的邻近搜索具体步骤如表8-1所示。

表8-1　基于平面扫描算法的邻近搜索具体步骤

步　骤	内　　容
1	对每个关键词$\text{Key}_i(i=1,\cdots,k)$的位置$p_{ij}(j=1,\cdots,n_i)$进行排序，生成位置列表
2	弹出每个位置列表最顶端的元素$p_{ij}(j=1,\cdots,n_i)$，根据它们的位置对k各关键词进行排序，找到最左和最右的关键词及其位置l_1和r_1，令$i=1$
3	从最左关键词的位置列表中找出顶端元素p，如果列表为空，转步骤6，如果$p>r_i$，那么区间$[l_i,r_i]$是最小化的，根据该区间大小r_i-l_i将其插入到一个堆中
4	在区间中移除最左关键词，同时从该关键字的位置列表中弹出顶端元素p
5	如果$[l_i,r_i]$是最小区间，令$r_{i+1}=p$，l_{i+1}为区间中第二个关键词的位置q，否则令$l_{i+1}=\min\{p,q\}$，更新区间和区间中关键字的顺序，令$i=i+1$，转步骤3
6	对堆中的区间进行排序，并输出结果

② 基于分治方法的k词邻近搜索。在基于平面扫描的算法中，需要将所有关键词的位置进行排序。然而，如果某个关键词出现的频率很低，则其他关键词的一些位置可以不进行排序而直接丢弃。因此，可以引入一种不需要排序的基于分治的搜索方法，步骤如表8-2所示。

表 8-2　基于分治的搜索步骤

步骤	内　　容
1	找到关键词的 n 个位置的中间位置 Mid
2	扫描位置列表，并将位置列表分为 Left 和 Right 两个列表，其中 Left 列表包含了比 Mid 大的位置，Right 包含了比 Mid 小的位置，保留 Left 中每个关键词的最大位置和 Right 中每个关键词的最小位置
3	使用平面扫描算法找到 Left 和 Right 之间的最小区间，这些区间用最后一步保留的位置所表示
4	如果列表 Left/Right 包含了所有 k 个关键词，那么递归地在 Left/Right 中寻找最小区间

3. 基于树与链表混合索引的检索方法

（1）模糊直方图模型。从每一帧音频数据计算归一化响度特征向量 $\boldsymbol{X}=(x_1,x_2,\cdots,x_{N_{\mathrm{FFT}}/2})$，其中 N_{FFT} 为傅里叶变换长度。若将该向量中的每个分量二元组表示为 $\langle f_k,x_k\rangle$，其中 f_k 表示 k 次谐波的频率值，x_k 表示 k 次谐波的归一化响度值，那么归一化响度向量可看成是二元组的集合。集合中的每个元素分布在不同的频率上，这样便可将集合中的元素映射到“频率-响度”二维平面上的一个点，为了叙述方便，称之为特征点。将该二维平面划分成为 N_{Ber} 个区域，每个区域和直方图的一个直方条（桶）相对应，并使用隶属度函数表示一个特征点属于某个直方条的程度。在音频数据的分析中，认为一个频率子带中的特征点对另外一个频率子带的隶属度为 0，而且一个频率子带内部的特征点与该子带中的一个响度值区间相关。这样一段音频信号 S 就可以用一个含 N_{Ber} 个直方条的模糊直方图表示为 $F(S)=[f_1,f_2,\cdots,f_{N_{\mathrm{Ber}}}]$，其中

$$f_1=\sum_{j=1}^{N_{\mathrm{Dot}}}\mu_j(i) \tag{8-14}$$

其中，N_{Dot} 为一段音频信号的特征点总数（等于帧数 $\times N_{\mathrm{FFT}/2}$）；$\mu_j(i)$ 为第 j 个特征点对第 i 个直方条的隶属度。

由于音频帧中大部分谐波分量的响度数值都比较小，只有少数谐波分量的响度数值相对突出一些，因此在直方图按数量统计响度数值时，这些最响的谐波分量虽较能体现数据帧间的差异，但由于数量较少而被其余的多数谐波分量所“淹没”，对直方图的贡献小，从而削弱了直方图对音频数据差异的分辨能力。为了增强直方图对音频数据差异的分辨能力，在直方图中只统计这些响度突出的谐波分量。

将所有频率分量分成两个集合，在直方图中再统计这些响度的谐波分量集合，这两个集合中元素的响度平均值差距明显，可按如表 8-3 所示的方法选择响度突出的分量。

表 8-3　响度突出分量的选择步骤

步骤	内　　容
1	初始化：设一帧音频数据的归一化响度特征向量为 $\boldsymbol{X}=(x_1,x_2,\cdots,x_{N_{\mathrm{FFT}/2}})$，集合 S 置为空集，表示选中的谐波分量集合，集合 $O=\{1,2,\cdots,N_{\mathrm{FFT}/2}\}$ 包含初始的全部 $N_{\mathrm{FFT}/2}$ 个谐波分量的编号
2	在集合 O 中，将响度最大的 m 个谐波分量 $k_2,k_2,\cdots,k_m$ 从集合 O 移到集合 S
3	在集合 O 中，将响度最大的一个谐波分量的编号选出，设为 k'，其响度为 $x_{k'}$，计算集合 O 中剩余谐波的响度均值 x_{avg}^{O}
4	考虑不等式：$$\frac{x_{k'}/x_{\mathrm{avg}}^{O}}{x_{\mathrm{avg}}^{S}/x_{k'}}=\frac{x_{k'}^{2}}{x_{\mathrm{avg}}^{S}x_{\mathrm{avg}}^{O}}>\lambda$$
5	其中，λ 为大于 1 的常数，如果上式成立则将 k' 加入集合 S，转 Step3；将归一化响度向量 $\boldsymbol{X}=(x_1,x_2,\cdots,x_{N_{\mathrm{FFT}/2}})$ 中不在集合中的谐波分量响度置零，算法结束

该算法可以自动根据一帧音频信号的“响度-频率”分布情况将响度突出的分量选出。

响度突出的分量不仅数量较少，在频域内的分布稀疏，而且数值差距也较小。由于纯音信号在频域上的掩蔽阈值下降较快，响度突出的分量间发生掩蔽效应的可能性也大大减小。因此，从降低算法复杂度和实际应用对计算精度的要求两个方面考虑，采用响度突出的分量并忽略掩蔽效应是合理的。

直方图之间的相似度度量方法有多种，其中直方图交集相似度方法是一种快速有效的度量方法。输入音频数据 I 与样例模板 R 之间的直方图交集相似度计算方法定于如下：

$$S(F^{R},F^{I})=\frac{1}{\min\left(\mathrm{sum}(F^{R}),\mathrm{sum}(F^{I})\right)}\sum_{i=1}^{N_{\mathrm{Ber}}}\min\left(f_i^{R},f_i^{I}\right) \tag{8-15}$$

其中，$F^{R}=[f_1^{R},f_2^{R},\cdots,f_{N_{\mathrm{Ber}}}^{R}]$ 为样例模板；$F^{I}=[f_1^{I},f_2^{I},\cdots,f_{N_{\mathrm{Ber}}}^{I}]$ 为输入模板；N_{Ber} 为直方图包含的直方条数量；$\mathrm{sum}(F^{2})$ 为对直方图的所有直方条数值求综合。这样定义的相似度可以反映不同的音频数据是否存在相互包含的关系。

（2）树与链表混合索引构造。对于任意三个直方图 F_1、F_2 和 F_3，直方图交集相似度具有如下性质：

$$S(F_1,F_2+F_3)\geqslant S(F_1,F_2) \tag{8-16}$$

其中，直方图做加法运算是将对应的直方数值相加作为结果直方图的数值。如果相似度阈值为 S_{th}，且长度不同的两段音频信号存在包含关系，则这两段视频的直方图相似度必定不小于 S_{th}。根据这个特点，可以基于直方图模型，采用链表与二叉树相结合的数

据结构为音频数据构造索引。

(3) 基于数与链表混合索引的搜索。由于直方图交集相似度的特点，在索引构造中，随着索引层次的增加，可能出现这样的情况：样例模板与某个非叶节点的相似度大于阈值 S_{th}，但该节点却不包含样例音频。如果这样情况的出现频率超过50%(使用二叉树结构)，则会降低使用索引的检索效率。从概率统计的角度看，这种情况的出现频率和两者实践长度的比值有关。假设两段音频信号的时间长度比值不大于 $2D_{max}$ 倍时，相似度数值能有效反映真实情况，即能根据相似度数值正确判断二者间是否有包含关系，称 D_{max} 为相似度最大允许深度，其数值与所采用的特征及匹配模型有关。那么，当在索引树的第 i 层搜索长度为 $2^i t_o$ 的音频段时，则可使用索引树 $i \sim i+D_{max}$ 层(与倍数 $2D_{max}$ 对应)间的节点进行快速搜索。可以试验证明输入模板长度分别是样例模板长度的2倍、4倍、8倍、16倍时的相似度变化情况。

检索时应该根据样例音频的长度在检索树中选择合适的层次范围来搜索。若样例音频的长度为 t_R，搜索层次下限 Lower 为

$$\text{Lower} = \lfloor \log_2(t_R/t_0) \rfloor \tag{8-17}$$

其时间粒度记为 $2^{\text{Lower}} t_0$，即在索引树中选择时间粒度不大于检索目标长度的最高层次。搜索的层次上限为

$$\text{Upper} = \text{Lower} + D_{max} \tag{8-18}$$

在将音频数据划分片段建立索引时，片段的边界可能和数据中包含的样例音频边界不重合。检索时使用直方窗从目标音频初始位置取出长度为 $2^{\text{Lower}} t_0$ 的音频数据建立样例模板，并在索引树中层 Lower～Upper 按深度优先遍历原则搜索样例模板。

(4) 时间复杂度分析。设检索源的数据长度为 t_s，样例的长度为 t_R，索引中叶节点的时间长度为 t_0，帧速率是 FPS(frames per second)，且样例在检索源中共出现 N_R 次。如果不采用索引结构，在检索源中用直方窗截取一段长度与样例相同的数据进行匹配，并逐帧向前移动，匹配的总数为 (t_s-t_R)FPS 次，时间复杂度为 $O(t_s-t_R)$。采用索引后，匹配次数约为

$$\begin{aligned} N_{\text{match}} &= 2\lceil t_s/(2^{\text{Upper}} t_0) \rceil + 2D_{max} N_R \\ &= 2\lceil t_s/2^{\lfloor \log_2(t_R/t_0) \rfloor} 2^{D}_{max} t_0) \rceil + 2D_{max} N_R \\ &\approx t_s/(2^{D}_{max}{}^{-1} t_R) + 2D_{max} N_R \end{aligned} \tag{8-19}$$

时间复杂度为 $O(t_s/(2^{D}_{max}{}^{-1} t_R)+2D_{max} N_R)$。若 t_s 较大、N_R 较小，则时间复杂度约为 $O(t_s/(2^{D-1}_{max} t_R))$。采用索引后，检索的速度将有大幅度提高，并且检索目标长度越大，检索速度越快，二者成正比关系。该索引方法的不足是采用直方图模型会导致存储开销大。

8.4.3 基于 GPU 通用计算的音频样例快速检索

面对海量的实时多媒体数据,检索速度一直是检索处理的关键问题之一。提高检索速度可以同时从两个方面进行:①从软件方面改进检索算法降低计算复杂度;②通过硬件平台提高检索过程的计算速度。NVIDIA 公司于 1999 年提出了图形处理器(graphic processing unit,GPU)的概念。GPU 在处理大量数据的并行计算方面明显优于中央处理器(central processing unit,CPU),目前已逐渐形成一个利用 GPU 进行通用计算的热潮。在生命科学、计算流体动力学、医疗成像等领域,已有很多研究成果出现。在音频处理研究领域,国外学者从 2007 年开始利用 GPU 对海量数据进行处理。

1. 通用图形处理器统一计算机设备框架

(1)通用图形处理器。随着芯片制造工艺的不断提高,GPU 拥有越来越强大的数据处理能力,如强大的并行处理能力和可编程流水线,可以处理非图形数据。基于 GPU 的通用计算是指用 GPU 来计算原本由中央处理器处理的通用计算任务,这些通用计算常常与图形处理没有任何关系。在单指令多数据(single instruction multiple data,SIMI)且数据处理的运算量远大于数据调度和传输的需要时,通用图形处理器在性能上大大超越了传统的中央处理器。

GPU 由图形处理单元和可编程处理单元两部分组成。传统 GPU 的可编程处理单元由定点着色单元和像素着色单元两种类型组成。它们分别用于处理 3D 图像中的集合图元操作和纹理滤波。由于这两种可编程单元的数量固定,传统的 GPU 体系架构无法很好地满足定点流水线和像素流水线之间的负载平衡,从而导致效率的降低。特斯拉(Tesla)架构的 GPU 使用统一着色单元执行定点着色程序和像素着色程序,当执行通用计算任务时统一着色单元又称为统一处理单元。由于传统 GPU 只能使用其中的可编程像素着色单位,而基于特斯拉架构的 GPU 可以使用全部的可编程处理单元,因此可以获得更高的执行效率。特斯拉架构的 GPU 由存储器系统和可扩展流处理器阵列(scalable streaming processor array,SPA)两部分组成,它们之间通过总线相连,并可以分别根据需求独立扩展。存储器系统由三部分组成:存储器控制器、固定功能的光栅操作单元和二级纹理缓存。其中,存储器控制器用于控制片外动态随机存储显存,光栅操作单位用于对存储器内的数据进行颜色和深度操作。可扩展流处理器阵列由若干线程处理器群组成,每个线程处理器群又由多个流多处理器组成。流多处理器由六部分组成:流处理器、特殊运算单元、多线程取值发射单元、指令缓存、只读常量缓存和读写共享存储器,它包含独立的完整前端,但一个线程处理器群中的所有流多处理器共享同一套存储器流水线。

(2) 统一计算设备框架。统一计算设备框架(compute unified device architecture，CUDA)是显卡厂商 NVIDIA 于 2007 年推出的不需借助图形学 API 就可以使用类 C 语言进行通用计算的开发环境和软件体系。CUDA 是与硬件无关的，程序经过一次编译后就可以在支持 CUDA 的不同规格 GPU 上运行。它被广泛应用于天文计算、生物计算、流体力学模拟、音频编解码、图像处理等诸多领域。

① CUDA 的运行方式。CUDA 的基本思想是将计算任务映射为大量的可并行执行的线程，程序执行时硬件会动态调度这些线程的运行，对并行度高的数据处理任务能有效发挥 GPU 的处理优势。在 CUDA 模型中，CPU 作为终端，而 GPU 作为写处理器运行一些能够被高度线程化的程序。

运行在 GPU 上的程序称为核函数(kernel)。核函数采用线程网格(grid)的组织形式，每个线程网格由多个线程块(block)组成。核函数的执行单位是线程块，各个线程块并行执行，彼此独立无法通信，没有执行顺序。线程网格中线程块的数量取决于问题的规模，而与具体硬件设备无关。同一个线程块内的线程可以彼此通信协同工作，这一特性显著提高了程序的执行效率。由于一个线程块中的线程需要共享数据，因此它们必须在同一个流多处理器中发射，线程块中的每个线程被发射到流处理器上执行。线程块和流多处理器是多对一的关系，即一个线程块被分到一个流多处理器，一个流多处理器在同一时刻可以有多个活动的线程块等待执行。这样，可以有效地隐藏时延，提高执行单元的利用率。

CUDA 采用单指令多线程的执行模型。单指令多线程是对单指令多数据执行模型的一种改进，两者的主要区别在于以下两方面：单指令多数据程序必须知道每条指令的宽度，向量的宽度受到硬件的限制，数据在打包成向量后才可以被处理。单指令多线程隐藏了 GPU 硬件 warp 指令的宽度，硬件能够自适应不同的执行宽度。一个线程块中线程]数可以在 1～512 取值，它们组成若干个线程束，每个线程束可通过一个 warp 指令执行。如果 CUDA 采用单指令多数据的执行模式，则每个线程块的宽度都必须与 warp 指令的宽度相应，这会很大程度上降低编程的灵活性。一个单指令多数据向量中的各个元素共享寄存器资源，不用考虑同步问题，向量之间的通信比较方便；单指令多线程中每个线程都有自己的私有寄存器，为了实现线程间通信，CUDA 引入了共享存储器和同步机制。

② CUDA 存储器体系模型。CUDA 的存储器模型中有六种存储器：寄存器、局部存储器、共享存储器、全局存储器、常量存储器和纹理存储器，其中各个存储器的特点如表 8-4 所示。

线程拥有自己的寄存器和局部存储器，线程块内的线程共用一块共享存储器，线程网格内的所有线程可以访问同一块全局存储器及只读存储器、纹理存储器和常量存储器。

表 8-4　CUDA 各个存储器的特点

存 储 器	特　　点
寄存器	GPU 片内的一种高速缓存，对于每个线程来说是私有的，访问延迟很低
局部存储器	对于每个线程也是私有的，访问延迟很大，如果寄存器被耗尽，数据将被存储在局部存储器中
共享存储器	一种 GPU 片内高速存储器，其读写速度几乎与寄存器一样快，同一个线程块内的所有线程都可以对其进行读写，它是线程间通信的最好方式
全局存储器	位于显存中，GPU 和 CPU 都可以对其进行读写，全局存储器能够提供很高的带宽，但其访问延时也很高，因此为了有效利用全局存储器必须严格遵守合并访问的要求
常量存储器	位于显存的一小块只读存储空间，适用于存储程序中频繁访问的只读参数，常量存储器具有缓存机制
纹理存储器	由 GPU 中用于纹理渲染的图形专用单元发展而来，它是一种只读存储器，最高可以存储三维数组形式的数据，它带有二级缓存机制，可以使用比常量存储器大得多的存储空间

2. GPU 音频检索加速算法

(1)检索算法可移植性。CUDA 程序优化的最终目的是以最短时间在允许的误差范围内完成给定的计算任务。“最短时间”是指整个程序的运行时间，更加侧重于计算的吞吐量，而不是单个数据的延迟。在开始使用 GPU 与 CPU 协同计算之前，应该先粗略评估一下使用 CUDA 是能够达到预想效果的。下面结合 CUDA 的特性，从精度、延迟和计算量三个方面来对音频检索算法进行分析。

从精度角度来看，目前采用 CUDA 的 GPU 无法满足高精度的计算需求，GPU 单精度计算性能远远超过双精度计算性能，整数乘法、除法、求模等运算的指令吞吐量也较为有限，即 GPU 最适合进行单精度浮点运算。而检索算法对数据的精度要求并不严格，采用单精度浮点数完全能够满足精度要求。

从延迟角度看，由于 CUDA 不能单独为某个处理核心分配任务，必须采用先缓冲一定的数据再交给 GPU 进行计算的工作方式。这种方式能够获得很高的数据吞吐量，然而单个数据经过缓冲、传输到 GPU 计算，再复制回内存的延迟就比直接由 CPU 进行串行处理要长很多，这就要求对实时性应用的要求不能很高。如果必须在数十微秒内完成对一个输入的处理，采用 GPU 可能会影响系统的整体性能，应该考虑现场可编程门阵列(field-programmable gate array，FPGA)或数字信号处理器(digital signal processor，DSP)来实现。检索子系统对延迟的要求在毫秒量级，而且系统可以通过增加缓冲大小来

进一步降低对延迟的要求。因此,从理论上看,系统能够容忍 CUDA 的延迟条件。

从计算量角度来看,如果计算量太小,那么使用 CUDA 是不划算的。因为在使用 CUDA 计算时,会因为访存和数据传输而增加时间开销。虽然 GPU 的单精度浮点处理能力和显存带宽都远远超过 CPU,但由于 GPU 使用 PCI-E 总线与主机连接,它的输入和输出的吞吐量受到了 I/O 带宽的限制。当计算密集度很低时,执行计算的时间远远比 I/O 花费的时间短,那么整个程序的"瓶颈"就会出现在 PCI-E 带宽上。此时,无论如何提高浮点处理能力和显存带宽,都无法提高系统性能。根据阿姆达尔定律可知,如果可以并行的部分在整个应用中所占的比例较低,那么 GPU 对程序整体性能的提高也不会非常明显。在分段检索中,尤其是样例模板较多时,每输入一个新的实时音频片段,都要到各个样例模板中进行滑动匹配,存在大量的距离(相似度)计算,计算量大正是系统的核心"瓶颈"。由于每次运算都是完全相同的操作,而且两次运算之间耦合度很低,原本串行的滑动匹配中的距离计算完全能够用并行计算的方式实现。

综合上述三点可以清楚看出,分段检索算法能够满足 CUDA 的限制条件,比较适合采用 CUDA 进行运算加速。

(2) 音频检索算法的计算特点。在考虑 GPU 加速方法时,应根据 CPU 和 GPU 的计算特点,充分挖掘两者的计算能力,从而最终达到整个系统的高效快速运行。基于分段的实时音频检索系统主要由三个步骤组成:样例模板加载(包括样例模板读入、特征提取等)、音频流的片段及特征提取和片段匹配,下面分别分析这三个步骤。

在音频检索系统中,为了方便用户辨别样例模板,通常使用音频格式的样例模板以便用户可以播放,这就需要在样例模板加载时计算模板特征,因此存在一定的计算开销。样例模板加载时间的长短对实时在线检索匹配的效率没有直接影响,但是,当样例模板数量较多时,完成一次样例模板加载需要的时间较长。例如,1 万个 20s 的样例模板完成一次加载大概需要 20min。随着系统处理能力的不断提升,模板库的规模会进一步加大,样例模板的加载速度也会变得越来越重要。一种解决的方法是使用样例的特征文件作为模板,或者在模板文件中同时保存音频数据域特征数据;另一种方法是对样例模板加载进行提速。

实时音频流的分段及特征提取是一个不间断的运算过程,主要有音频数据流输入就必须进行这一步骤,所以整体运算量会比较大。其中,音频流分段由 CPU 处理即可满足当前系统的需求,没有必要采用 GPU 加速;特征提取是一个相对耗时的过程,而且在样例模板库建立时也需要进行大量的特征提取操作。因此,可以对特征提取进行 GPU 加速,这样可以同时达到对前两个步骤进行加速的效果,而且样例模板越多,音频流越长,则加速的效果就越明显。

片段匹配实际上分为片段向量相似度计算和特征向量序列相似度计算两个过程，前者是输入片段向量在样例模板上滑动计算相似度，后者是输入片段特征向量序列在样例模板特定位置上计算向量序列相似度，这两个过程的共同点都是需要计算向量相似度。如果采用余弦距离作为相似度量，则两个过程的计算操作相对统一，而且都可以并行化实现。因此，可以相应编写两个核函数分别完成这两部分计算。

可以直观地说明音频检索算法中的三个步骤的计算量大小。例如，加载 10^4 个 20s 的样例模板，音频流输入长度为 12h，三个步骤所占比重如图 8-8 所示。其中，样例模板加载、音频流分段与特征提取、片段匹配分别占检索总时间的 17%、14%和 69%。因此，对片段匹配步骤进行加速能够获得最大的加速效果，应重点实现该步骤的 GPU 并行化计算。

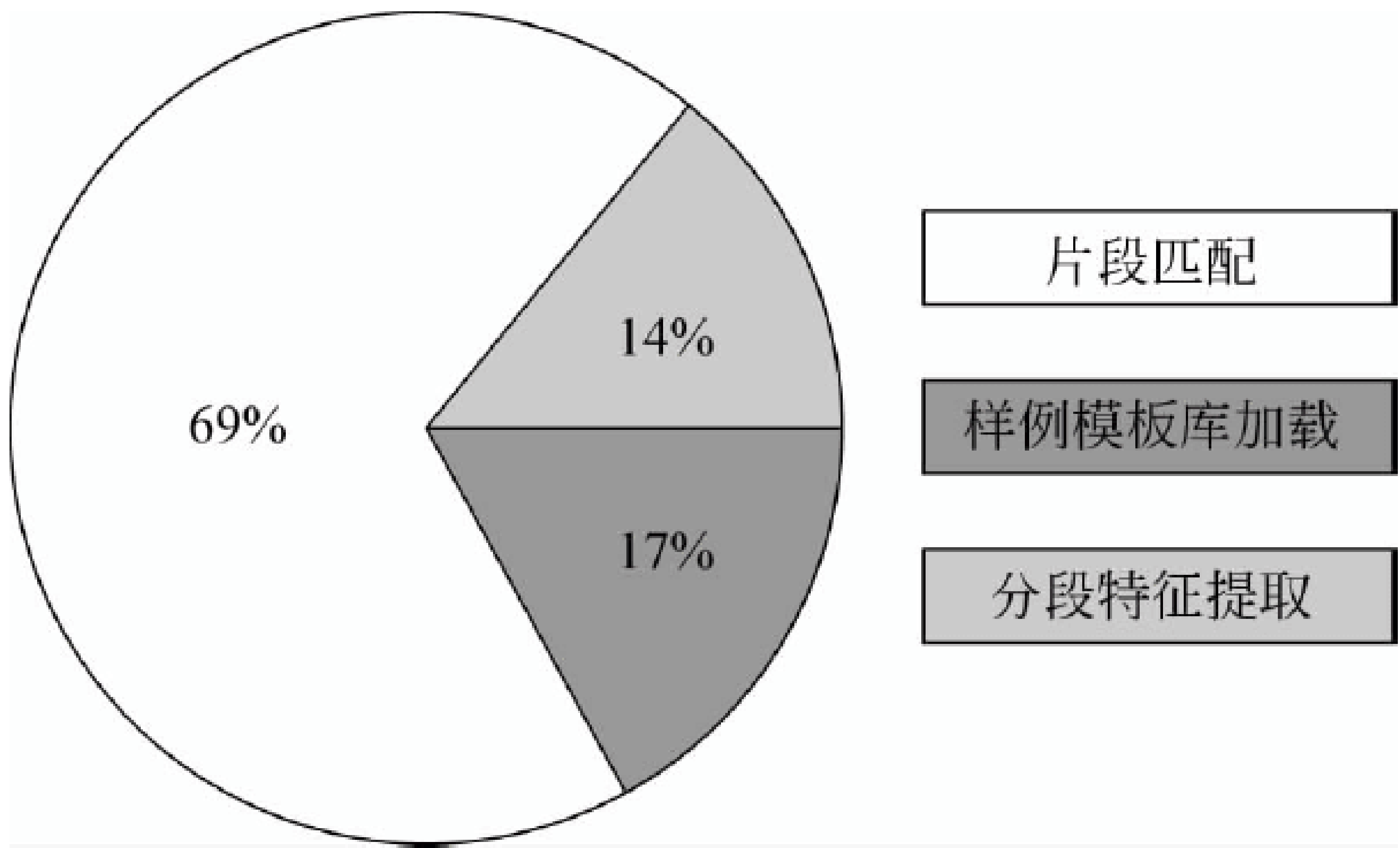


图 8-8　三种不同检索步骤的时间比重

可以将音频样例特征向量序列称为 Idx1 矩阵或 Idx0，将两个片段的音频样例特征向量序列匹配简称为 Idx0 匹配；将音频样例特征向量序列按照式(8-20)得到的量化值序列称为一维索引 Idx1，将输入片段 Idx1 值构成的片段向量在样例 Idx0 上的滑动匹配，即片段向量滑动匹配，简称为 Idx1 匹配。

$$p_i = \sum_{j=-\text{Range}}^{\text{Range}} c_j d(X_i, X_{i-j}), \quad \text{Range} > 0, \quad \sum_{j=-\text{Range}}^{\text{Range}} c_j = 1 \tag{8-20}$$

在基于分段的实时音频检索中，计算量比较大的运算操作包括：快速傅里叶变化、向量归一化、音频样例特征向量量化、Idx0 匹配、Idx1 匹配等运算。这样样例模板库大小为 10^4，样例模板长度为 20s，音频流输入长度为 12h 时，各个运算操作所占计算量比例见表 8-5。从表中可以清楚看出，检索的计算量主要集中在 Idx0 匹配、Idx1 匹配两个部分，它们占总计算量的 81.9%，是检索中的核心计算部分，因此应该是 GPU 提速的重点。

表 8-5　不同运算的计算量比例

运算名称	计算量比例/%	运算名称	计算量比例/%
Idx1 匹配	70.6	向量归一化	1.6
Idx0 匹配	11.3	MFCC 特征向量量化	1.2
快速傅里叶变换	2.4	其他	12.9

3. 基于 GPU 加速的音频检索算法应用

上面从三个不同的角度说明了将音频检索算法中的特征提取和片段匹配作为加速的对象。由于片段匹配所占比例最大，将其作为音频检索算法在 GPU 上移植的重点。根据并行化的程度及不同的存储空间使用，可采取以下两种实现方法。

(1) 以线程为粒度的方法。即一次片段匹配由一个线程完成，这种处理方式对音频检索算法改动较小，实现的重点是如何有效管理多种存储器、如何提高处理数据流及如何处理分配计算任务等。

(2) 以线程块为粒度的方法。即一次片段匹配由一个线程块完成。一个 CUDA 线程仅完成一次片段匹配中的一部分，一次片段匹配由一个线程块内的所有线程合作完成。为了适应 CUDA 编程的特点，对音频检索算法进行了一定程度的改进，在存储器管理策略和任务分配方式上区别于以线程为粒度的方法。

样例模板的数据结构：在音频检索系统中，样例模板中除了存放 Idx0、Idx1 外，还存放着其他不需要传输到 GPU 上的数据，如模板名称、类型等辅助描述信息。为了节省相对小的显存空间，仅需要样例模板的 Idx0 和 Idx1 传输到 GPU 显存中。因此，GPU 样例模板的主要数据结构如图 8-9 所示。

音频流片段组的数据结构：为了减少 GPU 的 I/O 交互次数，提高系统效率，应该一次性向 GPU 传输尽可能多的音频流片段。可以将多个片段组的主要数据结构设计为如图 8-10 所示，由于每个音频流片段等长，因此根据音频片段数量即可从音频流片段组的 Idx0、Idx1 中截取每个音频流片段的数据。

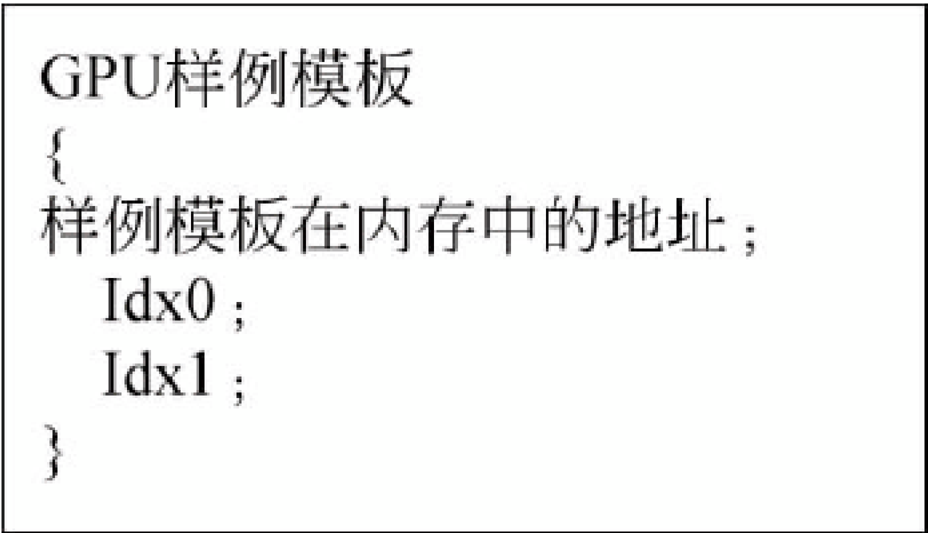


图 8-9　GPU 样例模板主要数据结构

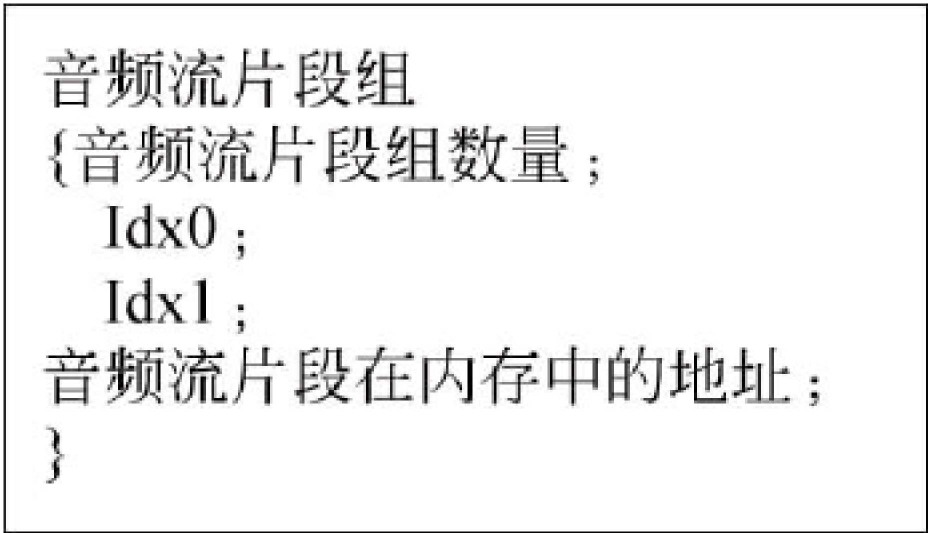


图 8-10　音频流片段组的主要数据结构

Idx1 匹配结果：Idx1 匹配结果使用的主要数据结构如图 8-11 所示，其中结果头单元用来记录该线程所处理的音频流片段在内存中的地址、该片段匹配的样例模板个数以及第一个结果单元的指针；结果单元用来记录与结果头单元中音频流片段匹配的样例模板指针以及在样例模板中的匹配位置。

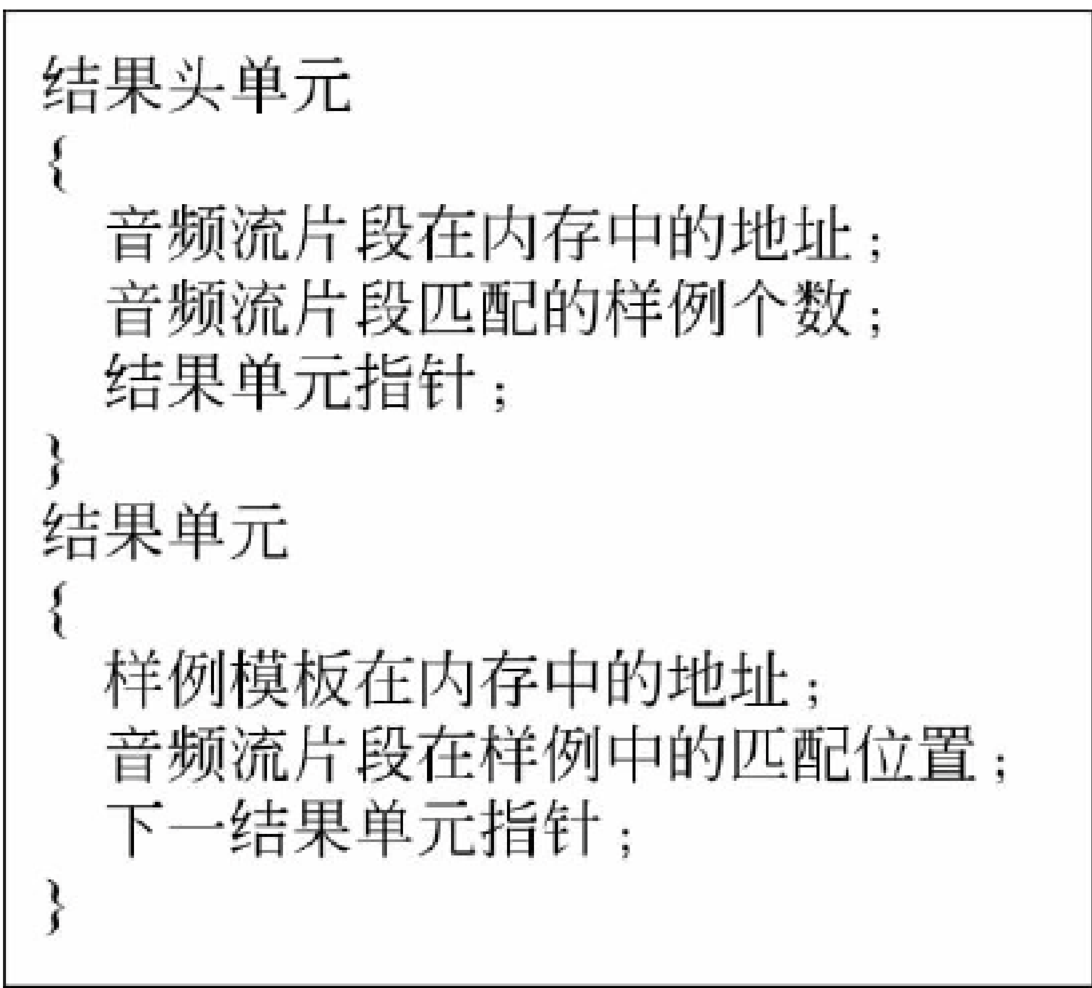

结果头单元
{
音频流片段在内存中的地址；
音频流片段匹配的样例个数；
结果单元指针；
}
结果单元
{
样例模板在内存中的地址；
音频流片段在样例中的匹配位置；
下一结果单元指针；
}

图 8-11　以线程为粒度时 Idx1 匹配结果的主要数据结构

将一组音频流片段传输到 GPU 后，Idx1 匹配结果将被保存在显存上一块连续的存储空间中。由于一个 CUDA 线程负责一个音频流片段的一次完整匹配，因此每个线程都必须分配自己的结果空间，并根据其线程号在结果空间中分配相应的位置，每个线程的结果空间分为两部分：结果头单元和结果单元。见图 8-12。

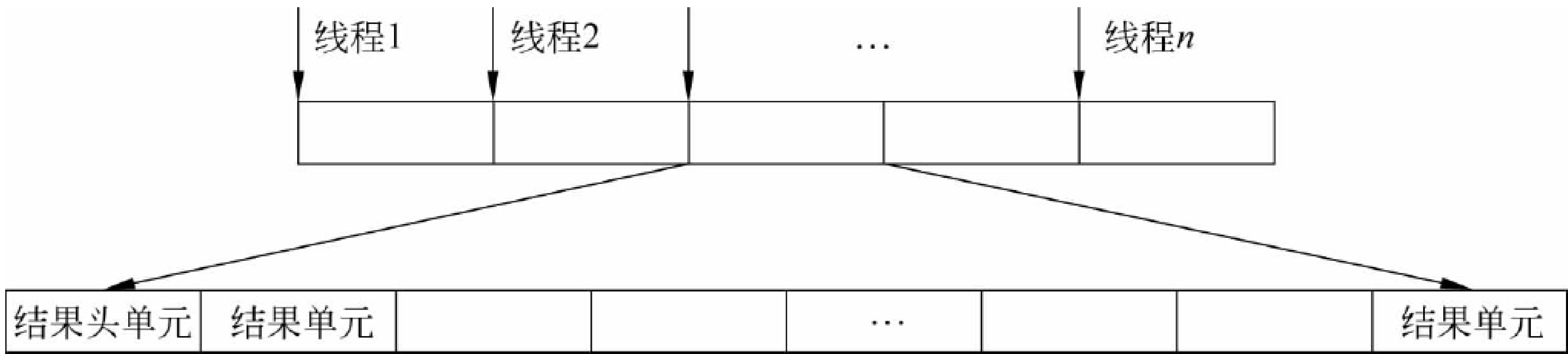


图 8-12　以线程为粒度时的 Idx1 匹配结果组织形式

Idx1 匹配节点：GPU 将 Idx1 匹配结果传回内存以后，遍历整个结果空间，将其中需要进行 Idx1 匹配的片段或模板信息完整地保存在数据结构 Idx1 匹配节点中，并组织成链表形式的队列，Idx1 匹配节点的主要数据结构如图 8-13 所示。

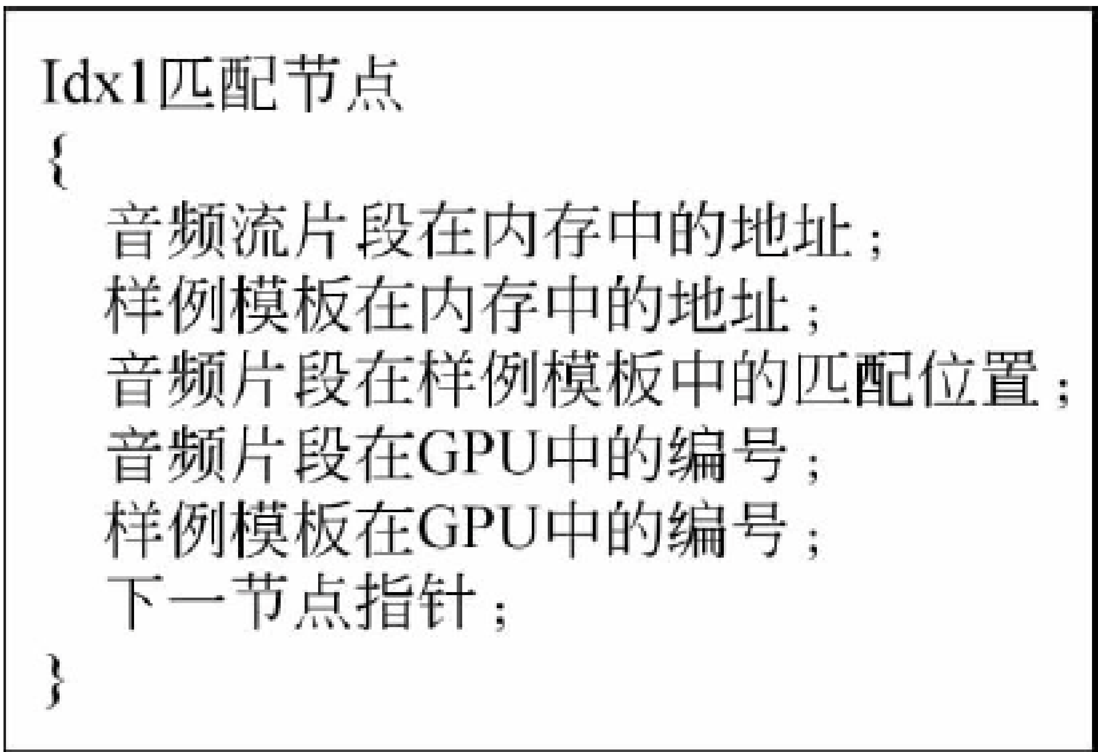

Idx1匹配节点
{
音频流片段在内存中的地址；
样例模板在内存中的地址；
音频片段在样例模板中的匹配位置；
音频片段在GPU中的编号；
样例模板在GPU中的编号；
下一节点指针；
}

图 8-13　Idx1 匹配重点的主要数据结构

系统根据 Idx1 匹配节点链表的节点信息进行 Idx1 匹配，若匹配成功，则认为数据流片段匹配是正确的。根据分段检索方法的原理，通过上下文信息最终获得检索匹配的完整结果。

8.5 语义级的语音文档检索

8.5.1 语音文档检索的预处理

语音文档检索(spoken document retrieval，SDR)有时也称为语音数据检索(spoken data retrieval)或语音检索(speech retrieval)，它是指为大量语音数据的内容构建索引，然后根据用户提出的查询请求，从索引中搜索和返回与用户请求相关联的语音文档或语音片段的处理过程。

21 世纪以来，随着多媒体技术的迅猛发展及其应用的日益广泛，越来越多的多媒体信息被人们记录并保存在计算机中。为了更高效地访问、管理和利用这些数据，人们迫切需要针对多媒体信息的检索方法。语音往往是多媒体信息不可或缺的重要组成部分，在多媒体信息检索任务中，语音文档检索扮演着非常重要的角色。语音是语言的载体，它在声学表示中富含可供检索利用的语义内容，而且它所蕴含的情感表示和韵律变化等特征又提供了高于语义层次更加丰富的信息。

目前，已应用的检索系统主要基于文档元数据(metadata)实现对语音与多媒体数据的检索。元数据一般是通过人工方法获取的音频文件内容的文字描述。这种方法虽然比较准确，但问题也很多：①标注多媒体数据需要大量的人工，而网络上却每天都涌现出海量规模的新的多媒体数据；②由于标注工作量繁重，往往标注内容仅能包含标题、关键字、内容简介等基本内容，因此可利用的索引资源非常有限；③对于较长的内容，没有办法提供查询词的时间定位和导航，给使用者带来很大不便。语音文档检索系统的框架结构如图 8-14 所示。为了实现快速检索，一般将检索任务分成“离线索引”和“在线检索”两个阶段来完成。

在“离线索引”阶段又包含“预处理”和“索引建立”两个处理环节。语音是不利于检索的声学信号，所以必须通过“预处理”环节将语音的声学表示级信号转化成更容易理解和处理的语义级信息。在已有的语音文档检索研究中，此“预处理”环节毫无例外是通过自动语音识别(automatic speech recognition，ASR)技术来实现的，通过 ASR 技术识别语音内容，将其转化为对应的文本表示。在语音文档检索系统中可供采用的 ASR 技术有两种：连续语音识别(continuous speech recognition，CSR)和关键字检出(keyword

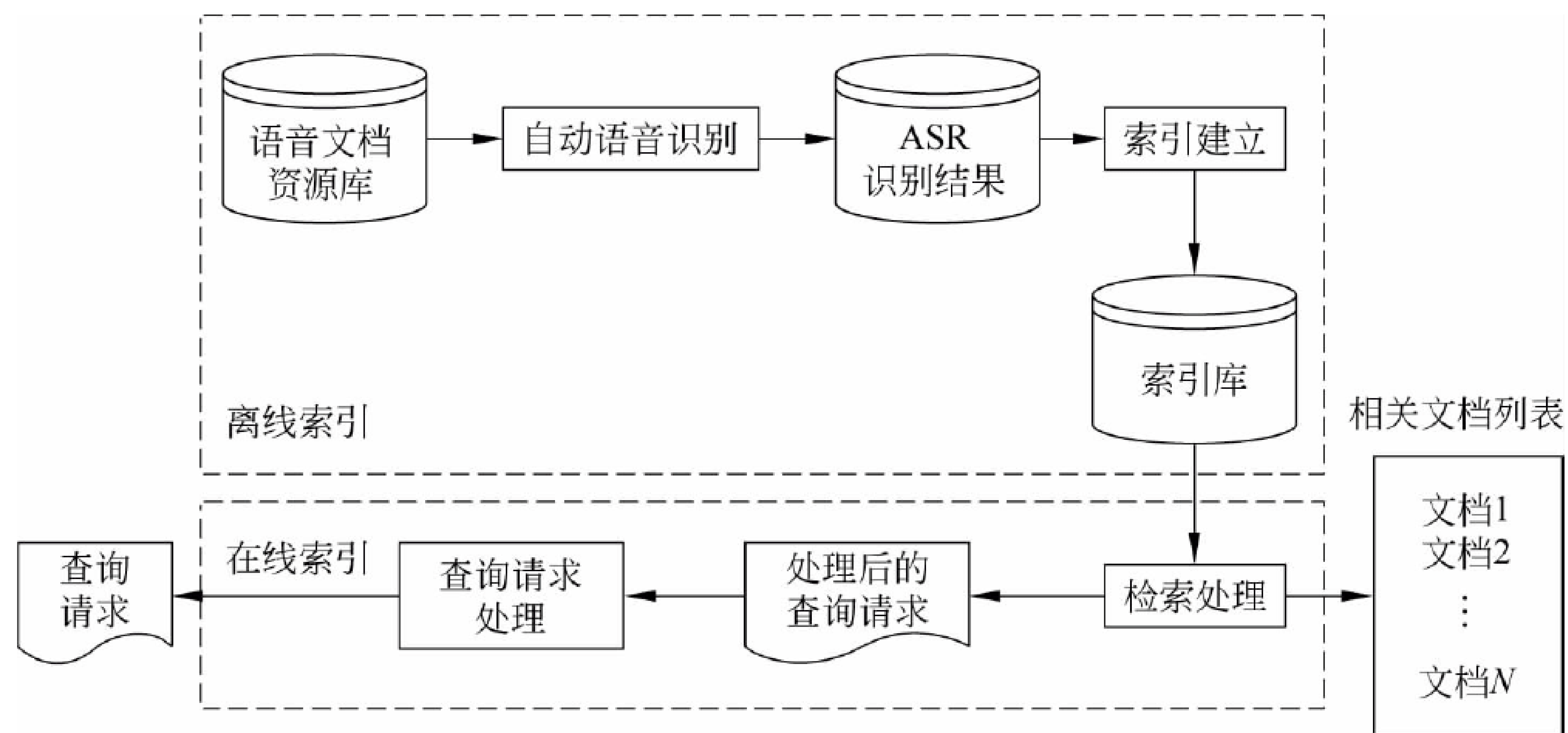


图 8-14 语音文档检索系统的结构

spotting,KWS)。连续语音识别技术识别整个语音内容,而关键词检出技术仅识别其中特定词汇。连续语音识别又有多种识别结果可供检索系统利用。

检索系统选择的 ASR 技术和其识别结果在形式上不同,语音的语义表示就不同,也就意味着检索系统在“索引建立”和“检索处理”等环节也必须采用不同的实现方法,因而可以根据目前所采用的 ASR 技术和识别结果对语音文档检索做大致的分类。

“索引建立”环节在预处理后得到的语义级信息中,提取可以有效支持检索任务的统计信息,并采用易于搜索的数据结构对其加以组织和存储,从而形成语音资源的索引库。在线检索时,检索处理模块根据查询请求在索引库中进行快速搜索,从而得到检索结果。“索引建立”和“检索处理”模板的设计属于信息检索领域的研究范畴,但又与传统的面向文本的信息检索技术有所不同。目前,在 ASR 的识别结果中无法完全避免识别错误的存在,有些条件下识别错误还可能相当严重,此外其处理对象也可能是网格这种结构复杂的多候选形式,这些问题使得语音文档检索系统中的“索引建立”和“检索处理”等技术有其独有的特点。

语音文档检索系统的检索结果可以有两种呈现方式,最常见的方式是将各语音文档按照与查询请求的相关程度进行排序,然后系统返回按此相关度排序的文档集合,此时查询请求可以是一个词,也可以是多个词,甚至可以是另一段语音文档。还有一种呈现方式则更关心在语音文档中每一个查询词都被检出,并且时间定位都准确,这种方式下不对语音文档进行排序,而是返回检出结果,此时的语音文档检索系统类似于传统的关键词检出

系统。对这两种检索方式,前者更多应用于面向大规模语音资源库的检索任务,如互联网的多媒体搜索引擎;后者通常应用于小规模资源库搜索和导航,如个人电脑中语音邮件的管理等。

英语语音文档检索研究是从 20 世纪 90 年代开始的,主要研究机构有剑桥大学、麻省理工学院、卡内基梅隆大学、微软公司、惠普公司等。英美等政府部门相继设立国家项目对该技术进行重点支持,比较知名的研究项目有 THISL 和 NGSW 等。由欧共体研究与技术支持的 THISL 项目(1997—2000 年),主要针对英式和美式广播电视新闻语言进行识别和检索研究。美国自然科学基金会支持的 NGSW 研究计划(2000—2005 年),旨在对美国国防高级研究计划局支持的负责各种语音处理技术性能评测的重要机构,于 1997—2000 年引入了语音文档检索专题,对语音文档检索系统的能力进行公平的评测。从 2006 年起,NIST 开始组织新一轮的针对大规模数据的查询词检索(spoken term detection,STD)评测,其应用目标是对以互联网为例的海量语音数据进行基于内容的检索、过滤和处理。NIST 组织的测评工作对全世界开放,极大地促进了语音文档检索的发展,使这一领域得到广泛关注。

针对汉语的研究则起步较晚,在汉语语音文档检索研究方面,中国台湾中研院资讯所、中国台湾大学语音实验室、中国台湾师范大学资讯工程系、中国香港中文大学人机通信实验室、微软亚洲研究院等学术机构开展了很多研究工作。早期的研究主要集中在如何针对汉语的特点来实现语音文档检索任务方面,近期的研究工作可以粗略地分为以下几个方面内容:跨语言的检索技术、概念层次的检索技术、面向 Web 搜索引擎的检索技术、混合索引技术等。哈尔滨工业大学语音处理研究室,在国家自然科学基金项目"基于音节网格的汉语语音检索技术"的资助下,开展了汉语语音检索研究工作,主要研究适合音节网格特点,并能够兼顾检索精度、索引尺寸、检索速度等各方面要求的检索方法。

语音文档检索有着非常广泛的应用领域。①对 Web 服务提供商而言,它是支撑多媒体信息搜索引擎的关键技术之一,用户可通过该搜索从互联网上快速获取所需要的多媒体资源;②通过该技术可以对广播电视、会议记录、语音邮件、讲座录音、有声读物等包含语言信息的多媒体文档,实现基于内容的检索、审查和有效管理;③可应用于情报搜集、信息内容安全等诸多领域,如监视非法的语音通信或管控网络中非法的音频流数据等;④可实现对数字图书馆中包含语言信息的资料进行分类和管理,并实现基于话题、谈话内容的检索;⑤信运营商可以通过该技术提供一系列具有高附加值的服务,如语言邮件管理、通话内容的实时记录和索引等;⑥语音文档检索技术的实现,使得政府机构和各种专业机构获得能够管理、分析和利用自己海量会议记录的有效手段;⑦对呼叫中心(call

center)等有着大量语音记录的服务机构而言,语音文档检索技术有着重要的使用价值;⑧个人电脑上的语音文档检索等。

8.5.2 语音文档检索的索引和搜索技术

当前主流的语音识别系统都采用基于统计建模的方法。设 $O=(o_1,o_2,\cdots,o_T)$ 表示语言声学观察对应的特征向量序列,令 $W=(\omega_1,\omega_2,\cdots,\omega_M)$ 表示对于声学观察的一个可能的词串。语音识别的目标就是在给定 O 的前提下寻找最可能的词串 W^* 基于贝叶斯决策理论,W^* 为最大化后验概率 $P(O|W)$ 的词串,即有

$$W^* = \arg\max_{W\in\psi} P(W \mid O) = \arg\max_{W\in\psi} \frac{P(O \mid W)P(W)}{P(O)} \tag{8-21}$$

其中,ω 为词串空间,由于语言识别系统所采用的词表是确定和有限的,所以 ω 也是确定和有限的;$P(O|W)$ 为在词串 W 时产生声学观察 O 的条件概率,通常基于声学模型计算;$P(W)$ 为词串 W 出现的先验概率,通常基于语言模型计算;$P(O)$ 为产生声学观察 O 的概率,因此在识别过程中一般不被考虑。语音识别系统最终应当选择使似然分数最大的 W 作为 O 的识别结果。

图 8-15 给出了标准的基于统计方法的语音识别系统的框架结构,包括前端处理、声学模型构建、语音模型构建、识别解码等过程。原始语言首先通过前端处理过程得到语言特征,在搜索解码过程中,通过语言特征与声学模型的匹配来计算声学分数,通常采用帧同步搜索的解码算法进行解码,最终得到识别结果。

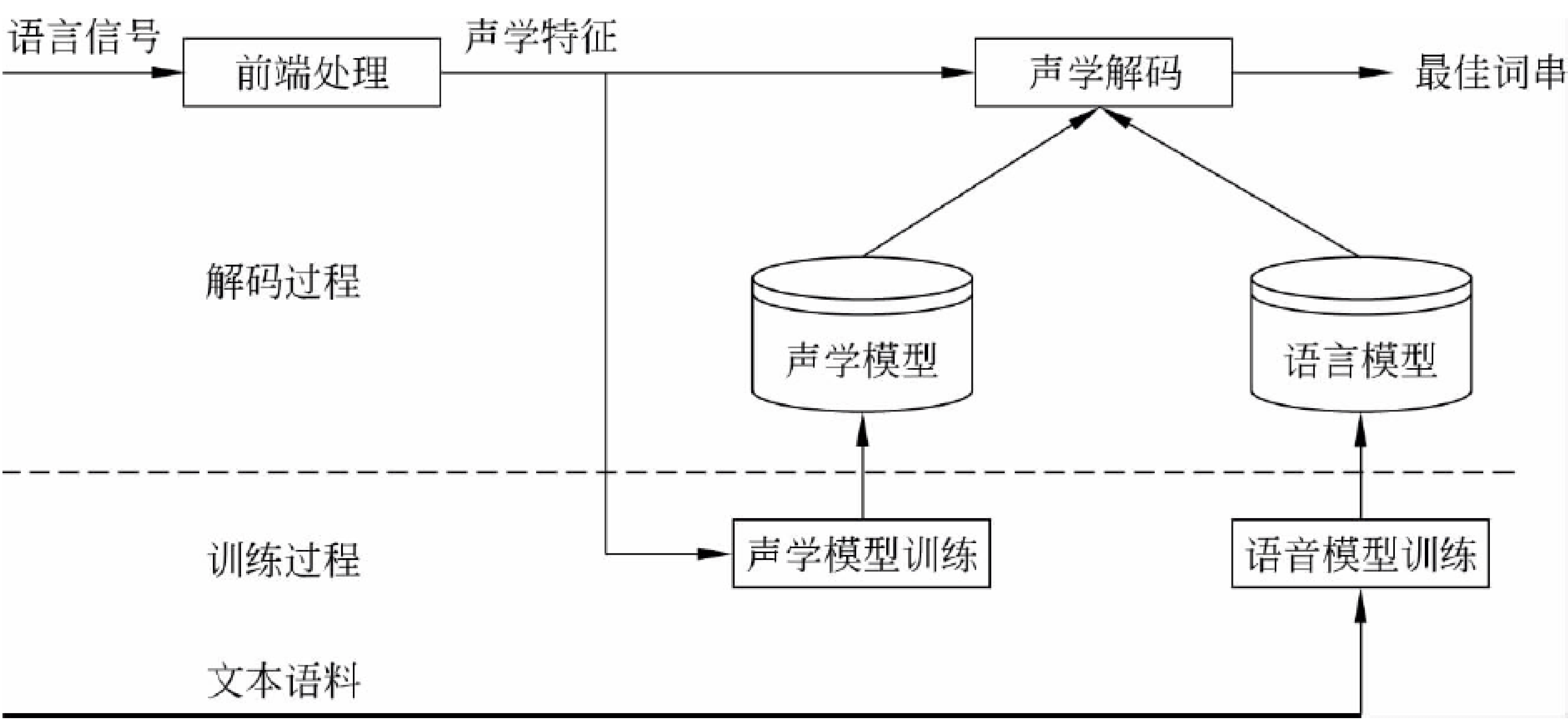


图 8-15 标准语音识别系统的基本框架

前端处理主要是对语音信号进行分析，完成特征提取、端点检测、去噪降噪处理等功能。特征提取的作用是使用更低维的向量表示原始语音信号。当前语音识别系统中最经常采用的特征有两种：线性预测倒谱系数 MFCC 和感知线性预测 PLP 系数，同时它们比其他特征应用也要广泛。

除基本特征外，也通常采用动态特征来刻画语音信号的时变特征，即原始特征的一阶和二阶差分或者其时间回归系数。语言识别器所采用的声学特征是由征组成的特征向量，如语音识别常用的 39 维 MFCC 特征，它包括 12 维 MFCC 特征、12 维 MFCC 的一阶差分和二阶差分、归一化对数能量、能量的一阶差分和二阶差分。去噪方面，最常见的是对特征向量进行基于整句语言的倒谱均值归一化和能量归一化等。

声学模型是用来描述特定语言单元声学特征的统计分布。声学建模基元数的选择依赖于具体的语言任务。通常来说，大的建模单元有更好的稳定性，但基元数目比较多，也不灵活；小的建模单元基元数目比较少，较为灵活，但稳定性比较差。对于大词表的语言识别任务而言，多采用音素作为建模基元，而对于中小词表的任务来说，通常采用音节甚至词作为建模基元。

(1) 隐马尔可夫模型。隐马尔可夫模型 HMM 是当前语言识别中最成功的声学建模技术，它能高效表征声学特征的统计特性和时变特性。隐马尔可夫过程是一个双重随机过程，其中之一是隐马尔可夫链，它描述了状态(非平稳信号的短时平稳段)如何转移到其他状态，另一个随机过程描述了状态和观察值之间的统计对应关系。由于站在观察者的角度，只能看到观察值不能看到状态，只能通过一个随机过程去感知状态的存在及其特征，因而称为“隐”马尔可夫模型。HMM 可定义为三元组 $H=(\pi,A,B)$，其中，π 为初始状态概率分部，A 为状态转移概率分部，B 为观察概率分布。π 和 A 刻画了语音信号产生的时变特征，B 则刻画了声学特征的统计特性。在语音识别中，通常采用无跨越自左向右拓扑结构的 HMM。每个状态上的观察概率分部则常采用高斯混合分部(Gaussian mixture distribution，GMD)形式。HMM 的训练是语言识别中的一个关键问题，通常基于最大似然估计(maximum likelihood estimation，MLE)准则，在大规模训练语料上估计 HMM 模型的参数。这一参数估计的过程可以采用最大期望(expectation maximization，EM)方法高效地实现，经典算法包括前后向算法(forward-backward algorithm，FBA)和 B-W(baum-welch)算法等。此外，为了提高模型区分能力，通常也采用其他有限准则来重新训练 HMM 模型参数，如最小分类错误准则等。

(2) 声学模型的训练。在连续语言识别中，通常采用音素作为建模基元。在语言中，协同发音(co-articulation pronunciation)现象非常普遍。当以音素作为建模单元时，为了

捕捉这种上下文不同导致的发音变化，通常采用两音素(bi-phone)或者三音素(tri-phone)来更精确地表示不同上下文的音素。这种扩展急剧增加了建模基元的数目，导致某种建模基元的训练数据不足。解决音频数据稀疏问题的主要方法是进行参数绑定，使某些模型参数可以共享同样的训练数据。混合绑定和状态绑定是两种常用的绑定方法。通常采用数据驱动或者决策树聚类的方法来进行参数的最优绑定。基于状态绑定上下文三音素的声学模型的训练流程图如图 8-16 所示。可分为四个步骤。

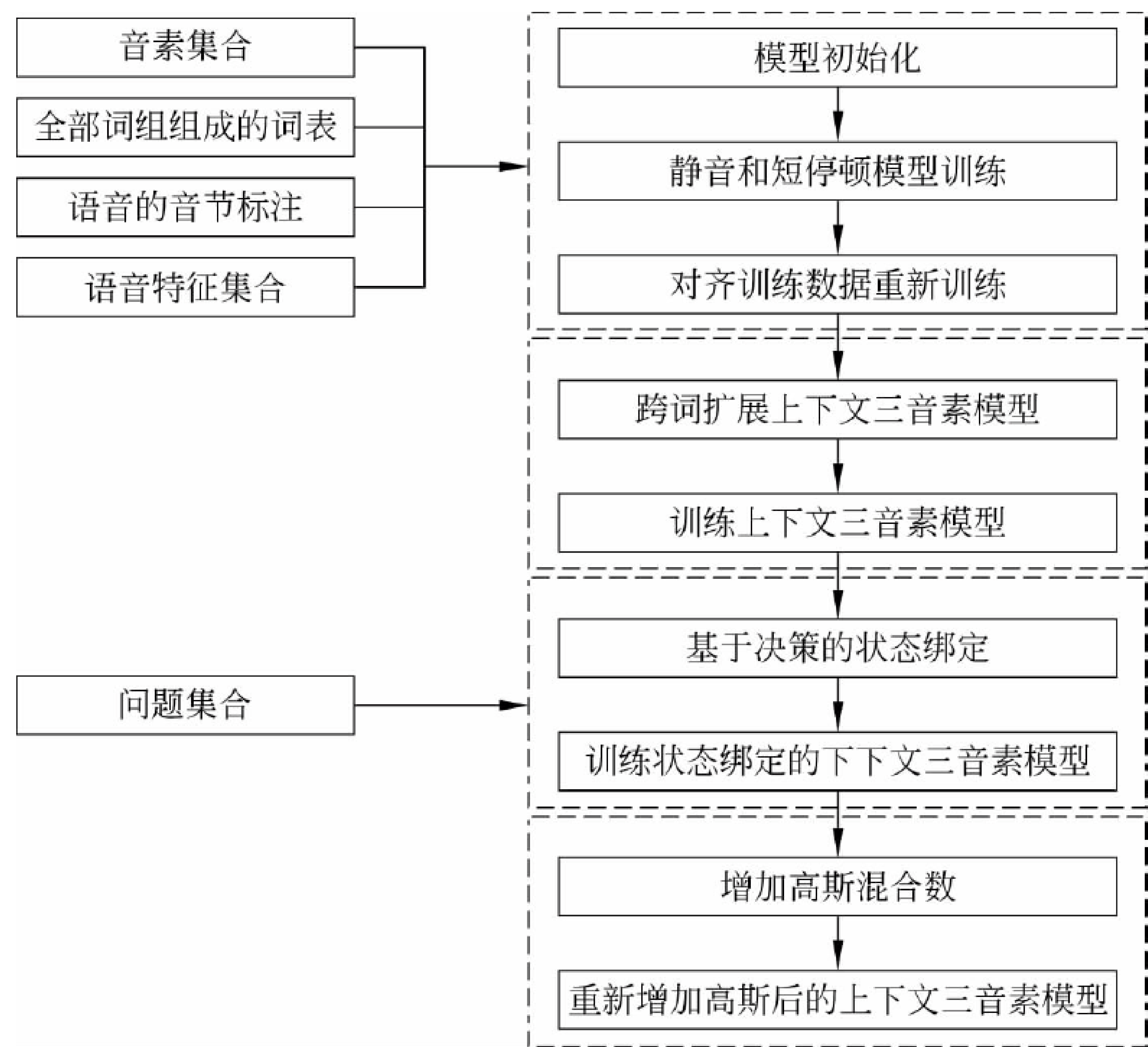


图 8-16　状态绑定上下文三音素模型的训练流程

① 单音素模型的训练。首先采用全局均值和方差作为全局初始模型，采用 B-W 算法训练上下文无关单音素模型的 HMM 模型；然后确定静音模型，并添加短停顿模型重新训练；最后对齐训练语料，获得语料的最佳标注，并重新训练模型。

② 跨词上下文三音素模型的状态绑定。把上下文无关单音素模型跨词扩展成为上下文三音素模型；然后重新训练扩展后的声学模型。

③ 上下文三音素模型的状态绑定。根据语言学知识，建立基于规则的问题集合，然后以数据和规则项结合的方式建立聚类决策树；把属于同一个叶子节点的上下文三音素

模型的状态进行绑定;状态绑定关系建立之后,再次重新训练模型。

④ 增加上下文三音素模型状态上的高斯混合数。增加每个状态下高斯混合数,并重新训练增加高斯混合数后的上下文三音素模型。此过程可以循环进行,直到状态上的高斯混合数达到给定数目。

(3) 汉语音素的建模方法。在汉语语音识别领域,最普遍的音素建模方法是音韵(initial-finals,IF)建模。汉语的音节由声母、韵母组成,一般声母仅包含一个辅音音素,而韵母则由一个原音或一个辅音加上一个原音组合而成。汉语是有调语言,声调信息是其区别于音域的一大特点。汉语中的每一个音节都对应一定的声调,共分五种:阴平、阳平、上声、去声和轻声,声调反映了说话人基频的变化趋势。采用声韵建模时,如音节"biang"可拆分为"b"和"iang"两个音素。音节后面及韵母后面的数字表示声调,其中 1～4 分别表示阴平、阳平、上声和去声,5 表示轻声。这样的音素基元总共有 187 个。

有调韵母分段模型(segmental tonal model,STM)是另外一种汉语音素建模方法。其建模方法如下。

① 对存在的韵母音节,将韵母位置前移与声母捆绑得到扩展声母集合。

② 用三个模型/H/、/L/、/M/对五个声调进行建模,将五声分别对应为/HH/、/LH/、/LL/、/HL/和/MM/。

③ 当韵母为双元音时,对前后的元音进行分隔建模。

表 8-6 所示为 STM 建模的示例,每一个音节都可以拆分成三个音素表示。

表 8-6　STM 音素建模示例

有调音节	音素 1	音素 2	音素 3
/huang1/	/hu/	/aaH/	/ngH/
/han2/	/h/	/aL/	/nnH/
/tiao3/	/ti/	/aaL/	/oL/
/da4/	/d/	/aH/	/aL/
/luo5/	/l/	/oM/	/uM/

(4) 关键词检出。关键词检出又被称为关键词识别(keyword recognition,KWR),是语音识别中的一个重要领域,其目的旨在从连续语音中检测并确认给定的若干个特定词。关键词检出与连续语音识别的主要区别在于,它不需要识别整个语音,而只需要识别其中感兴趣的词汇,识别时可以忽略语音中的其他内容。关键词检出算法结构如图 8-17

所示。

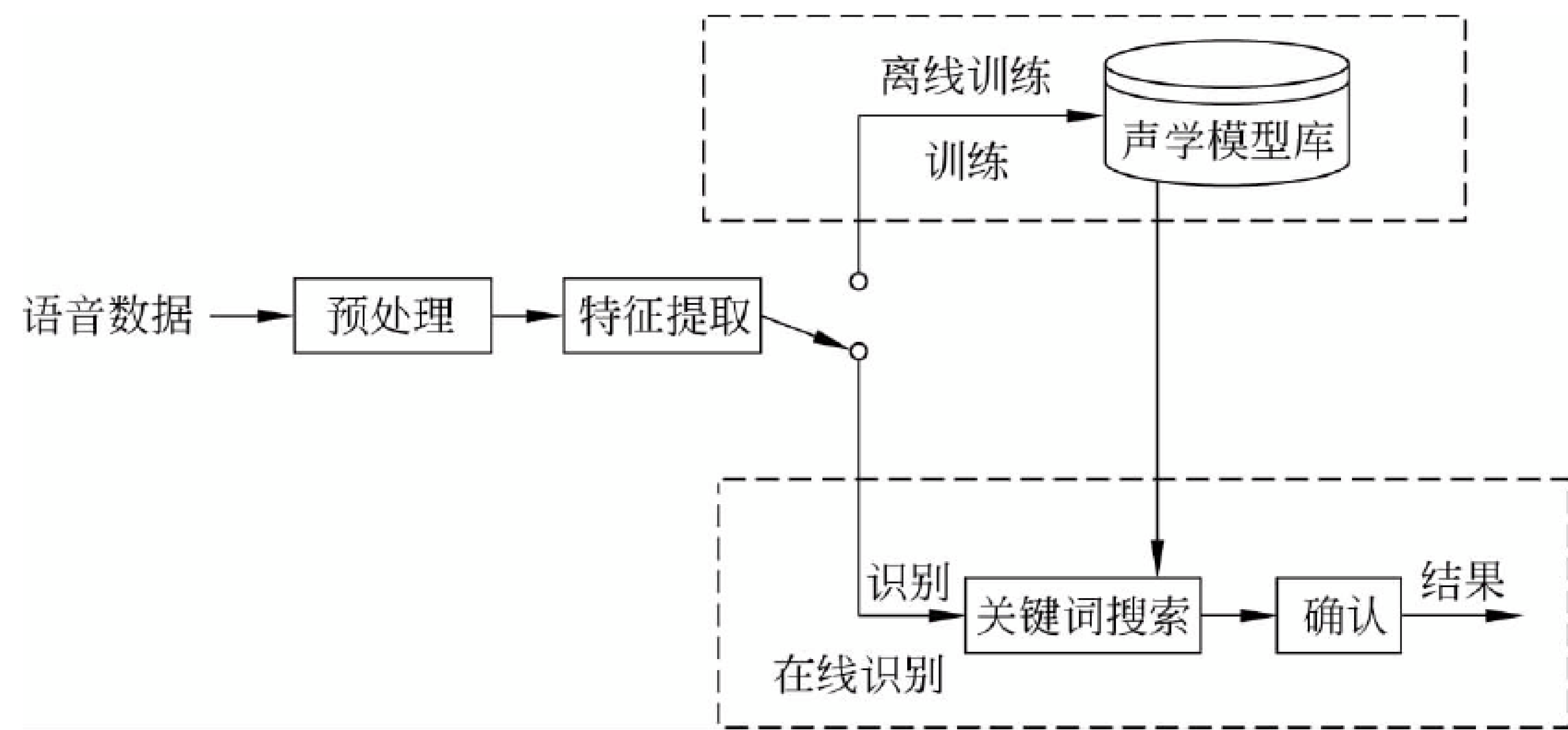


图 8-17 关键词检出系统基本框架

8.5.3 语音文档检索中的容错方法

语音文档检索是基于语音识别技术实现的,其检索性能受语音识别性能的制约。在这种条件下,提高语音文档检索系统容错能力是进一步提高检索性能的有效途径之一。所谓语音文档检索中的容错技术,是指能够在一定识别错误率下,通过提高检索系统对语音识别结果的容错性来提高整体检索性能。

研究容错技术的根本原因在于语音识别结果中缺失了一部分信息,或者说没有覆盖全部的正确内容。目前在语音文档检索中可采用的容错方法有以下三种。

(1) 采用模糊匹配的方式实现容错。虽然识别结果中有些信息不存在,但依据经验指导它们经常会被识别成其他特定的内容,那么当匹配到这些特定内容时,也可以认为缺失的信息以一定的可能性存在。

(2) 采用不同信息源相融合的方式实现容错。一种语言识别器会缺失这样的信息,另一种语言识别器会缺失那样的信息,但二者缺失的内容可能是不同的,如果检索时同时考虑不同识别器的识别结果,那么也许会取得更好的检索性能。

(3) 对识别结果进行修正和扩充。现有的语音识别技术无法完全排除表面识别错误的存在,但也许可以总结出特定识别器出现识别错误的规律,或者依据外部只是能够对那些识别错误做出准确的判断,从而可以通过纠错获得更加准确的识别内容。

1. 基于模糊匹配策略的容错方法

在基于子词最优候选的语音文档检索中,由于子词最优候选中识别错误比较多,提出

了许多降低匹配精度要求，采用模糊匹配的容错方法。采用模糊匹配的策略，首先要统计识别错误发生的先验知识，如麻省理工学院的研究者采用了音素识别错误混淆矩阵(phonetic recognition error confusion matrix)，矩阵中的元素 $C(r,h)$ 表示音素 r 被识别成音素 h 的次数，其中 r 是标注的音素标识，h 是识别结果中的音素标识。在进行文档相似度计算时，原有的方法是在隐身最优候选中搜索匹配表示查询词的音素串，即进行子串的精确匹配，若音素匹配成功，则匹配得分记为 1，否则记为 0，并累计匹配得分。当其等于音素串长度时可认为查询词匹配成功，从而得到语音文档中查询词的发生频次。采用模糊匹配策略后，音素匹配得分不再是 0/1 开关量，而是一个介于 0 和 1 之间的实数，音素 i 和音素 j 的匹配得分 $s(i,j)$ 可计算如下：

$$s(i,j) = \frac{C(i,j)}{C(i,i)} \tag{8-22}$$

2. 基于融合策略的容错方法

信息融合(information fusion)把来自多个信息源的数据和信息加以校准、联合、相关，合并成统一的表示形式以获得更加精确的信息。多源信息处理的概念并不陌生，它是人类和动物的一项基本功能，也是人类智慧活动的一部分。人们在从事生活、学习等各方面活动时，往往在综合考虑多方面因素后做出判断。信息融合是一个形式上的框架，在这个框架下通过融合的方式和工具将来自不同源的数据进行联合，从而达到获取质量更好的信息的目的。

近年来，在语音相关研究领域中，融合技术被广泛研究并采用，显著改变了识别系统的性能和鲁棒性。尤其是在说话人识别和语种识别中，融合技术更是成为不可或缺的技术手段。在语音文档检索领域也可以引入融合技术，通过使用不同的语音识别和检索技术构建多个检索系统，它们在特征、模型、检索方法等方面各不相同，具有较强的互补性，因此将它们融合能够获得更好的检索性能。一般而言，在越早的阶段进行信息融合，往往包含越多的信息，但是其需要构造的融合模型和处理算法也会相对复杂。考虑在模型层面和特征层面的融合比较困难，且结果层面的融合信息缺失较为严重，语音文档检索系统的融合主要是在索引和分数两个层面上进行的。

(1) 索引层面的融合。子词网格的结构，特别适合于索引层的融合。网格是一个“删减版”的解码网络，它是不同的特征和模型条件下解码网络中具有较高似然分路径的集合。不同来源的路径集合往往有很强的互补性，比单个网格包含更多的正确信息。基于融合网格的语音文档检索框架如图 8-18 所示。

在解码过程中，似然分数较小的路径被剪枝，只有似然分数较大的部分路径被保存并

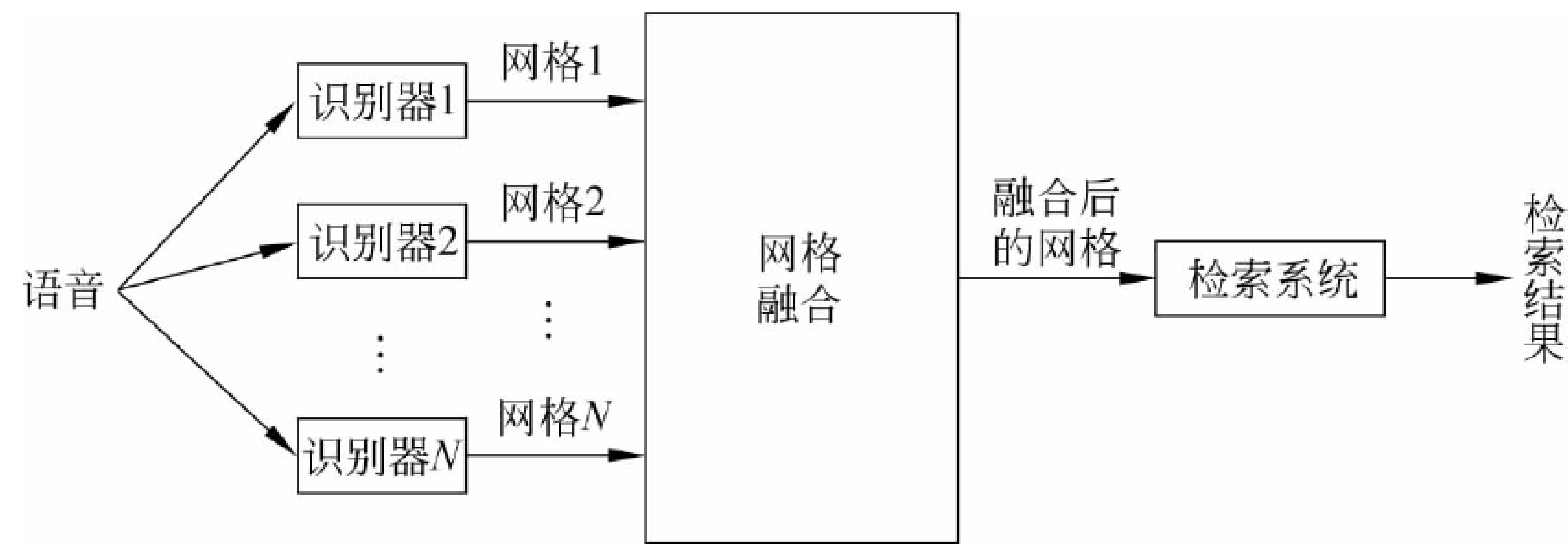


图 8-18　网格融合的系统结构示意图

写入网格。对网格进行融合，即对这些路径集合进行融合，需要考虑两个方面的问题：①拓扑结构的融合，如何将多个网格统一到一个网格中去；②分数的修正，如何计算融合后网格上匹配项的后验概率。

（2）分数层面的融合。网格融合提高了检索的整体性能，但该方法也有一定的局限性。这主要是由于该方法对参与融合各网格的结构有一定的限制，往往要求待融合的各网格采用相似的方法进行构建，并拥有相同的检索单元和后验概率估计方法。对在网络结构上不同构建方法的检索系统，不容易进行直接的网格融合。

分数融合是常用的一种融合方法。它将多个语音文档检索子系统输出的候选检索结果及相应的置信度分数进行融合。分数融合的系统框架如图 8-19 所示，目标语音分别进入不同的语音文档检索子系统，N 个子系统分别进行检索操作，得到检索结果——置信分数。最后对所有子系统输出的分数进行融合，得到最终的融合分数并进行判决。

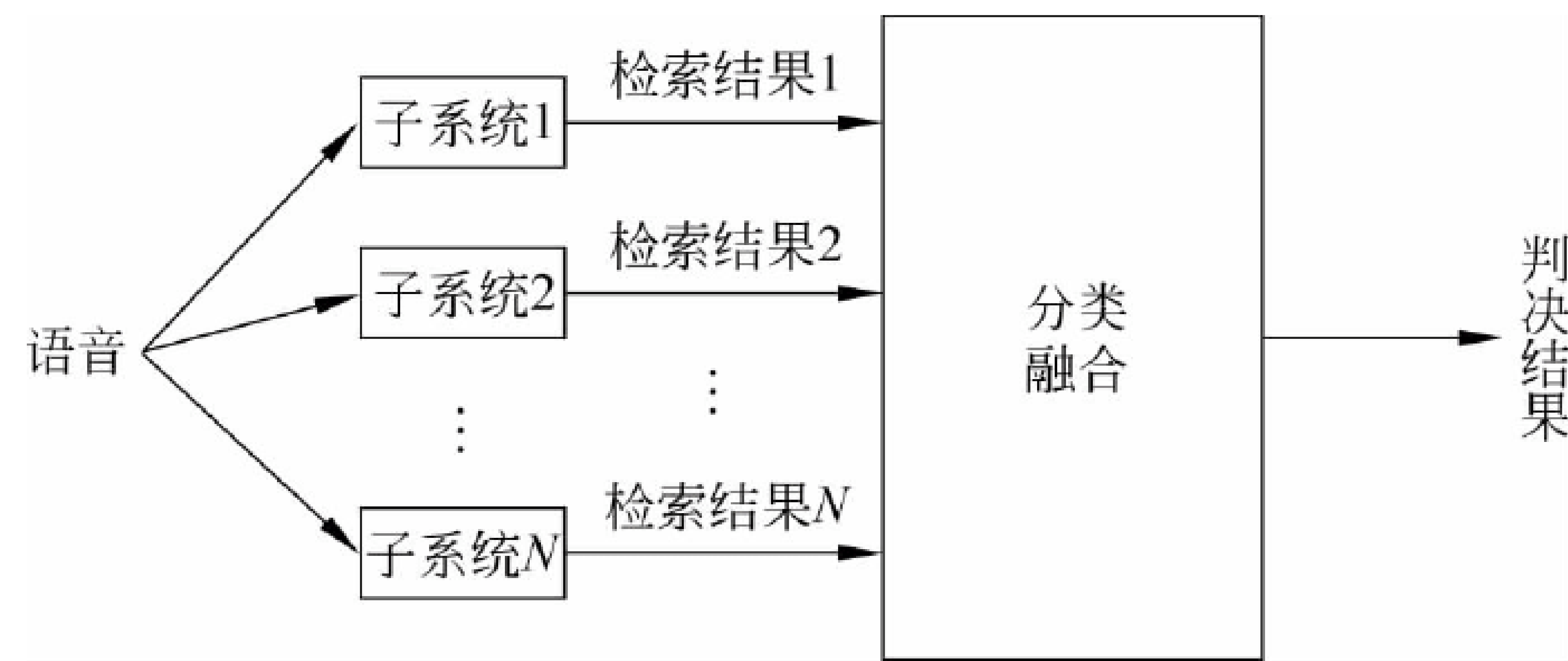


图 8-19　分数融合系统结构示意图

3. 基于扩充网络的容错方法

该方法是对基于音节网格的汉语语音文档检索研究任务提出的，用于对网格的内容进行修正，使得覆盖更多的正确内容。基于音节网格的汉语语音文档检索方法的检索精

度，依据非常接近在网络的最优候选上得到的检索精度。可以认为，它基本上达到了利用网格中最有候选得到检索结果的目的。一方面，这是一个令人鼓舞的研究成果，也充分说明了后验概率是一种非常有效的置信测度手段，它在语音识别所产生的网格结果中，能够起到区分正确信息和错误信息的效果；但另一方面，这也意味着受限于网格内容，很难再继续提高检索精度。导致这一性能“瓶颈”问题的根本原因不在检索方法本身，而是在于检索方法所给予的音节网格总是存在固有的错误率下界，即音节网格的准确性制约了检索方法可达到的最优检索精度。一般用网格错误率（lattice error rate，LER）来标识这个错误下界的位置。实验分析表明，当增加网格的多候选规模时，LER 下界的存在一定程度上反映了语音识别存在固有界限性能，或者说技术缺陷。而相应的检索实验表明，LER 的下降总能带来检索精度的提升。当 LER 达到稳定值时，检索精度也开始在一个很小的范围内波动，不再有大规模的改变。改善 LER 下界是提高检索精度的有效途径。

算法的基本思想：网格中错误下界的存在，往往意味着一些正确而有用的音节候选由于某种原因不能被包含在网格中，无论如何扩充网格的候选规模，它们都会被识别器遗漏掉。如果能够找到这些被遗漏的音节，以及它们被遗漏的位置，就能够将它们补充到网格中去，从而得到错误率下界被改善了的扩充网格。可以认为，被遗漏的音节与网格的数据分布间应该有一定的关联性，存在着某种统计规律。如果能够为此规律建立统计模型，就有可能根据网格的内容估计出被遗漏的音节。

4. 基于词片语言模型的容错方法

基于词片语言模型的容错方法，通过在语音识别器中引入新的子词基元，以达到扩充网格规模以包含更多正确内容，进而提高检索精度的目的，这就是基于音节网格的汉语语音文档检索容错方法。

在针对英语的语音文档检索中，研究者曾提出一种被称为“词片（word fragment）”的子词形式，它可被理解为经常重复出现的一组音素组合。在检索系统中利用这种子词基元能够有效改善检索性能。针对汉语的特点，减少有调音节合并成新的基元“词片”的方法。该方法基于互信息最大准则，利用迭代算法在文本语料中自动生成若干大于音节、小于词的词片基元。通过构造基于词片基元的语言模型，利用音节或音节之间相互搭配的语言学信息，从而使语音识别结果能够更好地体现词级的语言学信息，达到提高语音识别性能，同时降低网格的错误率下界的目的。

本章小结

需要检索的音频资源主要指能够被计算机处理的数字化音频(digital audio),它将在时间上和幅度上都是连续的模拟声音信号经过采样和分层处理,进行编码后得到离散数字表示的数字信号。音频内容从整体上看可以划分为三个等级:最底层的物理样本级、中间层的声学特征级和最高层的语义级。

音频检索处理方法可分为三类:一是语音检索,即以语音为中心的检索,采用语音识别等处理技术,例如电台节目、电话交谈、会议录音等;二是音乐检索,即以音乐为中心的检索,利用音乐的音符和旋律等音乐特性来检索,例如检索乐器、声乐作品等;三是音频检索,即以波形声音为对象的检索,这里的音频可以是汽车发动机、雨声、鸟叫等各种声音,也可以是语音和音乐等,这些声音都统一用声学特征来检索。

从检索技术及其依据的基本原理出发,音频检索仍然分为基于文本的音频检索技术和基于内容的音频检索技术两类。基于内容的音频检索(content-based audio retrieval)就是通过从音频数据中提取和分析音频特征信息,对不同音频数据赋予不同的语义,使具有相应语义的音频在听觉上保持相似。本章也主要是阐述基于内容的音频检索技术。

音频信息检索模型,就是在对音频信息进行抽象表达的基础上,通过构建一种评测机制能衡量用户查询请求与待检音频信息的相似度,即提供一种衡量用户查询请求与音频数据相似性的方法。通常可采取两者之间的距离或相似度概率来体现它们之间的相似性程度。目前的音频信息检索技术,其模型很大程度上借鉴了文本信息检索模型的思想。典型的模型包括向量空间模型和概率模型。

音频样例检索既可以应用于检索静态音频数据库,也可以应用于检索实时音频流。相对而言,检索实时音频流难度更大、要求更高,算法需要更多地考虑资源开销和计算速度问题。

基于 MPEG-1 压缩域模糊分类的音频检索方法是采用一种基于距离的模糊分类法,用隶属度刻画音频片段与类别之间的联系,认为每个音频片段与各个类别中心都有一个隶属关系,对不同类别之间有交叉的数据进行有效分类。

从高维空间的角度来看,检索过程就是给定任意一个查询点(向量),在数据库中找到与查询接近的点,并能保证以较高的概率返回与查询最接近的点。如果数据库的规模很大或数据的维数很高时,穷举法的实践代价往往无法接受。高维数据库的索引存在“维数的诅咒”问题,即索引的复杂度随维数的增加呈指数增长。

统一计算设备框架(compute unified device architecture,CUDA)的基本思想是将计算任务映射为大量的可并行执行的线程,程序执行时硬件会动态调度这些线程的运行,对并行度高的数据处理任务能有效发挥 GPU 的处理优势。CUDA 程序优化的最终目的是以最短时间在允许的误差范围内完成给定的计算任务。基于分段的实时音频检索系统主要由三个步骤组成:样例模板加载(包括样例模板读入、特征提取等)、音频流的片段及特征提取和片段匹配。

语音文档检索(spoken document retrieval,SDR)有时也称为语音数据检索(spoken data retrieval)或语音检索(speech retrieval),它是指为大量语音数据的内容构建索引,然后根据用户提出的查询请求,从索引中搜索和返回与用户请求相关联的语音文档或语音片段落的处理过程。语音文档检索技术属于基于语义的音频信息检索技术的研究范畴。

声学模型是用来描述特定语言单元声学特征的统计分布,隐马尔可夫模型 HMM 是当前语音识别中最成功的声学建模技术,它能高效表征声学特征的统计特性和时变特性。隐马尔可夫过程是一个双重随机过程。在连续语音识别中,通常采用音素作为建模基元。

当以音素作为建模单元时,为了捕捉这种上下文不同导致的发音变化,通常采用两音素(bi-phone)或者三音素(tri-phone)来更精确地表示不同上下文的音素。解决音频数据稀疏问题的主要方法是进行参数绑定,使某些模型参数可以共享同样的训练数据。混合绑定和状态绑定是两种常用的绑定方法。通常采用数据驱动或者决策树聚类的方法来进行参数的最优绑定。

语音文档检索中的容错技术是指能够在一定识别错误率下,通过提高检索系统对语音识别结果的容错性来提高整体检索性能的技术。目前在语音文档检索中可采用的容错方法有三种:①采用模糊匹配的方式实现容错;②采用不同信息源相融合的方式实现容错;③对识别结果进行修正和扩充。

本章思考与练习题

1. 简述数字化音频的含义。
2. 音频信息有哪些基本特征?
3. 简述音频信息的内容层次。
4. 音频信息检索技术可以分为哪几类?
5. 基于文本的音频检索方法有哪些突出缺点?
6. “基于内容的音频检索”的含义?

7. 音频信息检索模型的含义?
8. 向量空间检索模型和概率检索模型的基本含义?
9. 实时音频检索的主要技术难度是什么?
10. 简述基于 MPEG-1 压缩域模糊分类的流媒体音频检索方案。
11. 邻近搜索的含义?
12. 说明敏感哈希索引方法的一般原理。
13. 在 k 词邻近搜索中有哪两种算法? 各自的含义是什么?
14. 简述基于树与链表混合索引的音频检索方法。
15. 通用图形处理器的含义与作用是什么?
16. 简述 CUDA 的运行方式。
17. 描述基于平面扫描算法的邻近搜索有哪些具体步骤。
18. 描述基于分治方法的 k 词邻近搜索有哪些步骤。
19. 说明响度突出分量的选择步骤有哪些。
20. 统一计算设备框架(CUDA)有哪六种存储器,各自有何特点?
21. CUDA 程序优化的最终目的是什么?
22. 基于分段的实时音频检索系统主要由哪三个步骤组成?
23. 说明语音数据检索(或语音文档检索)的概念含义是什么?
24. 描述语音文档检索系统的结构。
25. 语音文档检索有哪些主要的应用领域?
26. 阐述基于统计方法的语言识别系统的基本框架。
27. 简述隐马尔可夫模型的含义。
28. 基于状态绑定上下文三音素的声学模型的训练流程有哪些步骤?
29. 在汉语语音识别领域,最普遍的音素建模方法是什么? 举例说明其含义。
30. 语音文档检索中的容错技术的含义与作用是什么? 可采用哪些容错方法?

第 9 章　视频信息检索

视频信息检索可以广泛应用于工业、农业、商业、科研与多媒体服务业等领域。例如，应用于大型监视系统中可以检测和搜索特殊类型的视频内容事件，实现监控系统的智能化；应用于电影电视行业，可以作为非线性编辑系统的一个组成模块，提供对大量的视频数据的组织和检索；应用于交互多媒体系统、数字图书馆或视频服务业中，可以为用户提供友好的视频浏览和视频交互检索界面，使用户更快地找到需要的视频信息。由于视频数据与其他数据在形式、结构、内涵等方面都不同，与图像相比，视频的结构更为复杂，数据量也更大，对基于内容的视频检索的要求也就更高。

9.1　数字视频的相关基础知识

1. 数字视频的基本概念

根据人眼的视觉停留特性，通常当画面显示速度超过每秒 25 帧时，人眼会将快速变换的画面视为连续画面，视频就是利用这样的原理来模拟真实动态世界的。

视频不像图像那样“一目了然”，人们使用视频的目的是从中获取信息，想要“一目了然”地了解一段视频中是否包含他们感兴趣的内容，这就需要将视频数据进行结构化处理。对于视频数据，至少有两个基本的层次结构：整个视频序列和单个的视频帧。但对于视频数据库的管理和检索来说，仅仅基于整个视频流的结构，就不能深入到视频内容，也无法实现基于内容的分析和检索；另一方面，由于视频庞大的数据量，如果是基于视频帧的处理，则运算量是相当大的，而且帧作为组成视频的最小单位，包含的信息量不大，用户也很少对视频中的单帧感兴趣。一般来说，一段视频由一些描述独立故事单元的场景构成，一个场景由一些语义相关的镜头组成，而每个镜头是由一些连续的帧构成，它可由一个或多个关键帧表示，其结构如图 9-1 所示。

因此有必要构造中间级的结构层次，构造便于检索的视频结构。本章中主要涉及的数字视频基本概念有以下四个。

(1) 帧(frame)。帧是视频流的基本组成单元，每一帧就是一幅图像。视频流就是由

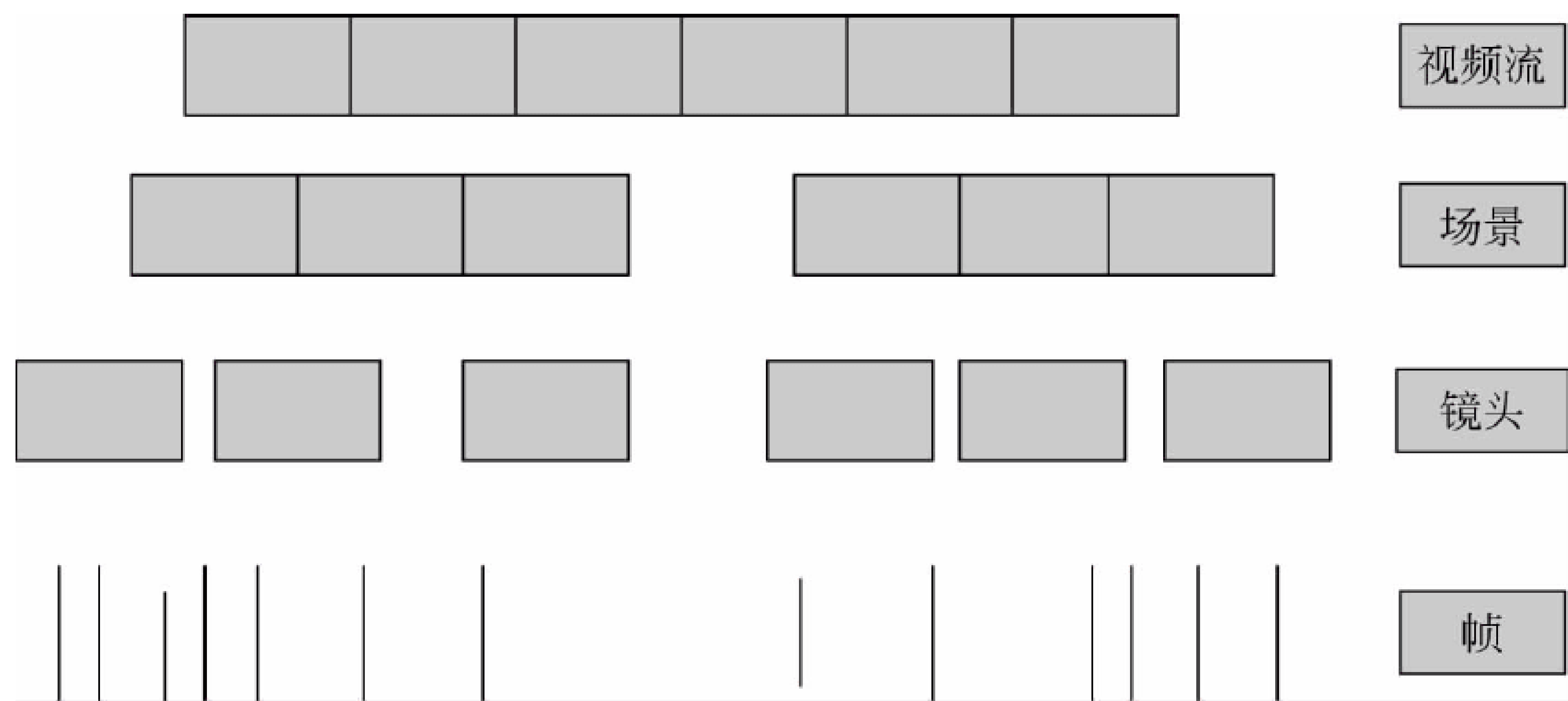


图 9-1　视频分层结构图

连续图像帧构成的。在 PAL 制式的视频中，帧速率一般为 25 帧/秒；在 NTSC 制式中，帧速率一般为 30 帧/秒。

（2）镜头（shot）。镜头是指由摄像机不间断拍摄的一组帧序列，它常被看成是视频的最小结构单元。一般来说，同一个镜头中的图像帧比较相似，其对应特征基本保持不变。因此，通过发现相邻帧之间较剧烈的特征变化，可以判断是否发生了镜头转换。

（3）关键帧（key frame）。关键帧有时也称为代表帧，用以描述一个镜头的关键图像帧，它可以用来代表一个镜头的主要内容。关键帧的使用大大减少了视频索引的数据量，同时也为视频摘要和检索提供了一个组织框架。

（4）场景（scene）。场景是由语义上相关、时间上相邻的若干镜头组成，它们一般发生在相同的时间和地点，出现相同的人物或事件。场景反映了视频所蕴含的较高层语义内容，如学校运动会这个场景可以由运动员入场、运动员比赛和观众呐喊等多个镜头组成，形成一个比较完整的语义表达。

2. 数字视频模型

视频数据库系统既包含了视频数据本身的内容，也包含了不同视频数据间的关联数据。视频数据库系统的基础是视频数据模型，数据模型包括数据结构和操作。其中数据结构既要研究与数据本身内容相关的对象，也要研究描述不同视频数据间关系的对象。而数据操作则只是对数据的各种加工利用手段，如对数据的插入、删除、查询等。数据模型有很多种，下面简单介绍两种模型。

（1）实体-关系模型。实体-关系模型是一种典型的数据模型，它包含以下几种基本元素。

① 实体(entity)。实体是客观存在的,既可以是真实的事物也可以是抽象概念。在视频数据库中,视频段、镜头、视频流以及对视频的注释等实际对象和概念都是实体。

② 标识符(label)。用来标识实体实例的名称。在视频数据库里,视频节目的名称就是一种标识符。

③ 属性(attribute)。属性指实体的特征或特性。在视频数据库中,对一个给定的视频段,可以用其内容,如其中出现的人数、人名、时间等作为属性来描述。

④ 关系(relation)。指实体间的关系。在视频数据库中,不同实体间可能有完全不同的关系。

(2) 语义对象模型。语义对象模型也是一种典型的数据模型,它比实体-关系模型更接近用户的感觉。该模型包含以下几种基本元素。

① 语义对象。语义对象是足以描述一个确切本题属性的命名集合,与实体-关系模型中的实体对应。

② 标识符。用来表示语义对象的名称,这种标识符是语义对象的潜在名字。

③ 属性。指语义对象的特性或特征,一般一个属性用作标识符就需要有值。在视频数据库中,对一个指定的视频段,可以用其内容,如其中出现的人数、人名、时间等作为属性来描述。

④ 属性域。是关于属性的可能取值的描述,域的特征依赖于属性的类型。

由于视频有其独特的性质,仅用传统的数据模型不能有效表达,为此要建立专用的视频数据模型。目前已建立了多种视频数据模型:时间线模型、时间层次模型、代数模型、视频对象数据模型。下面介绍一下视频对象数据模型。

在面向对象的视频信息数据库系统 OOVID(object-oriented video information database)中定义了一种视频对象数据模型。在一个视频节目中,任何一部分都可成为一个独立的视频对象,它有自己的属性和属性值。一个视频对象可有任意的属性和属性值,但它所独自具有的属性和属性值可以表达它所包含的所有视频帧序列的内容含义。视频对象是对有意义场景的描述数据,包括对象标号、时间间隔、一组属性-属性值。一个视频对象可以用一个三元组$[O,I,V]$表示,其中 O 是视频对象标号,I 是时间间隔集合的一个子集合,V 是一个 n 元组$[a_1:v_1,a_2:v_2,\cdots,a_n:v_n]$,其中每个 $a_n(1\leqslant i\leqslant n)$是属性名集合 A 中的一个属性名,v_i 的值可以定义为

$$v_1,v_2,\cdots,v_n(1\leqslant i\leqslant n),\quad \{v_1,v_2,\cdots,v_n\}$$

每个元素 $x\in D$(基本元素值集合)都是一个值;每个间隔 $i\in J$(时间间隔集合)都是一个值;$a_i:v_i$ 是一个值,称为集合值;每个视频对象也是一个值。

3. 数字视频的特点

数据视频内容丰富,结构复杂,不同于传统的字符型文本数据。它主要有以下几个特点:

(1) 视频数据量大。视频数据通常是利用图像采集设备将各个图像帧自动输入计算机而最终形成的,它不是结构化数据,而是以数字图像或数字视频的非格式化形式表示。从数据量上来看,一幅分辨率为640×480,颜色为24bit/pixel的图像数据量大约为1MB,如每秒播放30帧,则1秒钟视频的数据量大约为30MB,即使经过压缩,一部普通长度的影片也将占用数百兆空间,这显然绝非结构化记录数据所能比拟的。

(2) 视频数据结构复杂。文本数据是字符数值型数据,不含空间和时间属性,可以看作是一维数据。图像数据是一种具有空间属性的数据,但没有时间属性,可看做是二维数据。而视频数据不但具有空间属性还具有时间属性,是三维数据。空间维是每一个视频帧图像具有的空间结构,时间维是指视频是一系列沿时间轴顺序分布的视频帧形成的流结构。因此视频数据具有时空特性,从而视频数据的表达和模型的建立变得困难。

(3) 视频数据具有很大的冗余性。冗余性是指一个镜头的连续视频在一段时间内仅发生微小的变化,大部分数据是冗余的,这也是视频压缩的理论基础。

(4) 视频信息的丰富内容带来解释的多样性和模糊性。人们在观看一段视频时,对视频内容的理解往往加入了一定的主观因素,因此不同的人可能会有不同的理解,这就不像字符型数据那样只有一个客观的完全确切的解释。视频数据解释的模糊性,使得用户在进行查询时,无法像字符型数据那样用指定的关键字精确查询一个特定的记录,在视频数据库中,往往只能用相似性匹配的方法进行检索。

9.2 基于内容的视频检索系统结构

基于内容的视频检索(content based video retrieval,CBVR)指根据视频的内容及上下文关系,对大规模视频数据库中的视频数据进行检索。主要特点是直接从视频数据中提取信息线索,它是一种近似匹配,在没人工参与的情况下自动提取并描述视频的特征和内容。

基于内容的视频检索系统结构如图9-2所示。先将视频流通过镜头边界检测分割为镜头,并在镜头内选取关键帧,再提取镜头的运动特征和关键帧的视觉特征,作为一种检索机制存入视频数据库,最后根据用户提交的查询,按一定特征进行视频检索,将检索结

果按相似度呈现给用户,用户可以优化查询结果,系统会依用户意见灵活优化检索结果。特征的提取和检索算法的优劣决定了其效率和性能。

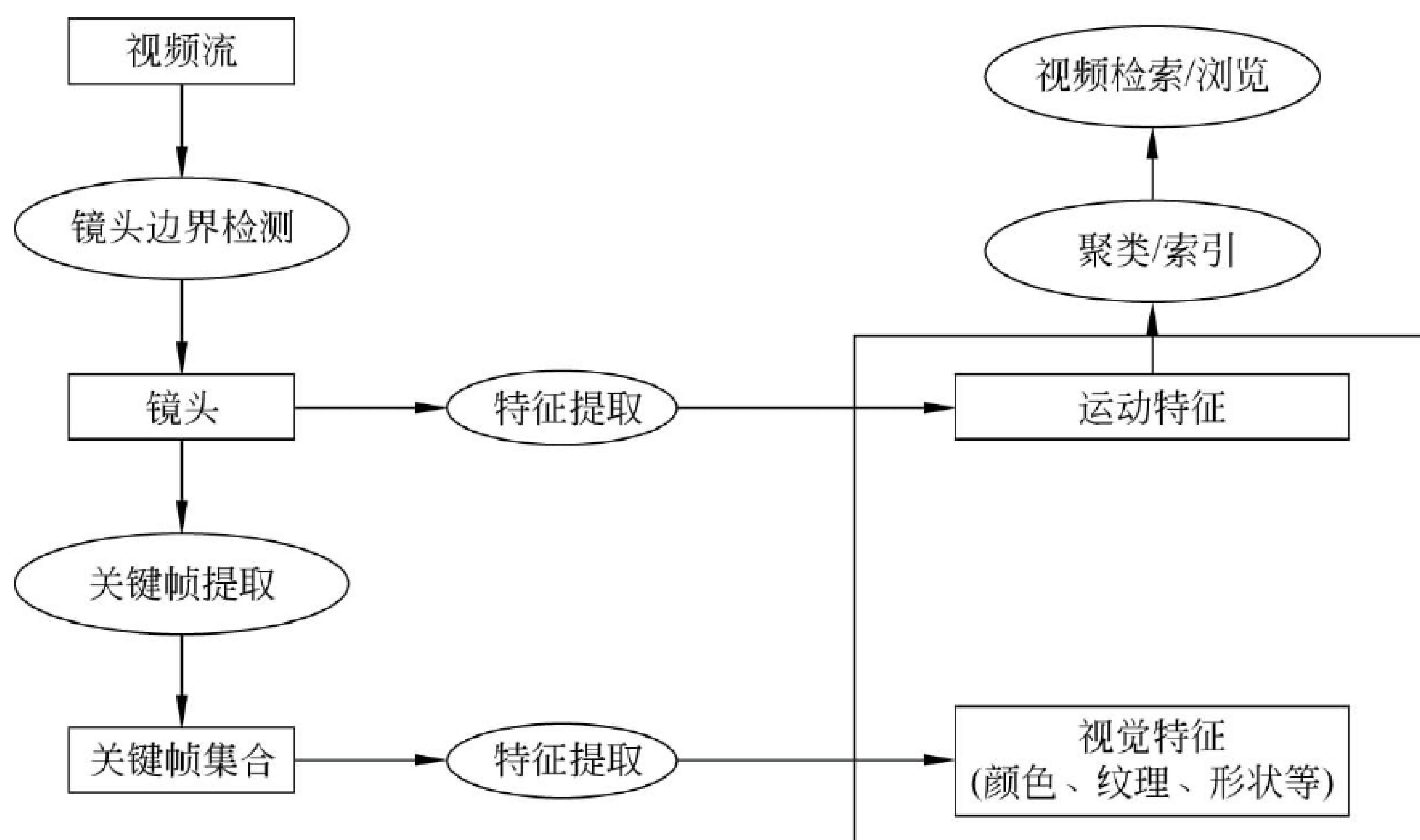


图 9-2　基于内容的视频检索系统结构图

9.3　视频镜头分割

镜头是视频数据的基本单元,所以基于内容检索的视频处理,首先要把视频自动地分割为镜头,以作为基本的索引单元,这个过程就称为镜头边界的检测,也叫场景转换检测(scene change detection,SCD),它是实现基于内容视频检索的第一步。

通常的边缘检测方法是先通过边缘检测算子找到图像中可能的边缘点,再把这些点连接起来形成封闭的边界。图像边缘提取不仅可以剔除不相关的信息,保留图像重要的结构属性,而且通过边缘提取可以使得信息处理量大大降低,从而降低整个算法的运算量,并且可以使其在抗噪性能上大大提高。

基本的镜头边界检测算法有两类:一类是基于图像特征的非压缩域边界检测,另一类为基于编码信息的压缩域边界检测。

非压缩域的镜头分割方法指先解压视频中的 I、B、P 各帧,然后通过计算图像间的特征差异检测镜头边界,如图 9-3 所示。这种方法可以得到比较高的检测精度,但是特征的计算量比较大,其中最典型最基本的有基于像素、直方图、块、边缘等方法。

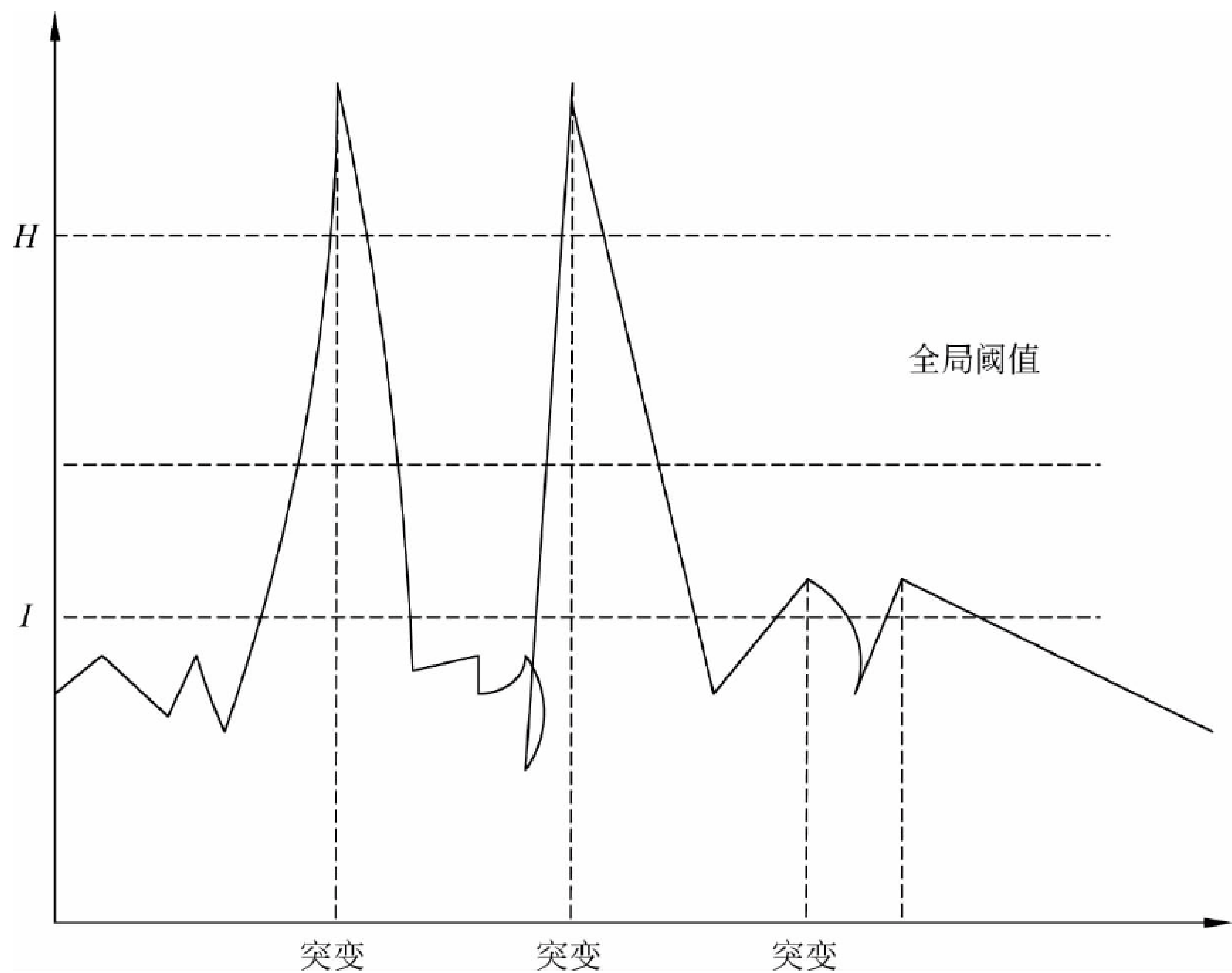


图 9-3 镜头分割帧差

压缩域的镜头分割方法指通过解析压缩域中的编码信息检测镜头边界，这些编码信息有 DCT 系数、帧间预测、运动矢量和宏块编码等。由于只需少量的解码即可获得这些信息，因此该方法的检测效率很高，常用于实时的镜头边界检测，其中最基本的有基于 DCT 系数、基于 DC 系数、基于运动矢量和宏块预测信息等方法。

9.3.1 非压缩域的镜头分割方法

1. 基于像素的镜头分割方法

由于最直接反映视觉内容的元素就是每个像素的灰度或亮度值，因此度量相邻帧之间差异最简单的方法就是计算第 k 帧和第 $k+1$ 帧中所有像素的灰度或亮度的差值绝对值之和，然后通过统计该总差值占总像素数的百分比来确定是否发生了镜头改变。于是第 k 和 $k+1$ 帧的帧间差 $Z(k,k+1)$ 可表示为

$$Z(k,k+1)=\frac{1}{XY}\sum_{x=1}^{X}\sum_{y=1}^{Y}\mid I_k(x,y)-I_{k+1}(x,y)\mid \tag{9-1}$$

其中：X、Y、$I_k(x,y)$、$I_{k+1}(x,y)$分别是图像的宽度、高度、第 k 帧中(x,y)像素的灰

度值、第 $k+1$ 帧中(x,y)像素的灰度值。这种方法的一个缺点就是不能区分大区域内的小变化和小区域内的大变化，这通常会造成镜头的无效检测。为区分上述两种变化，对此种算法的一种改进就是只计算灰度变化达到一定阈值 T 的像素个数，即帧间差 $Z(k,k+1)$可表示为

$$Z(k,k+1)=\frac{1}{XY}\sum_{x=1}^{X}\sum_{y=1}^{Y}D_{k,k+1}(x,y) \tag{9-2}$$

其中

$$D_{k,k+1}(x,y)=\begin{cases}1, & \text{if } |I_k(x,y)-I_{k+1}(x,y)|>T\\ 0, & \text{else}\end{cases} \tag{9-3}$$

式(9-3)中，1 表示该像素发生了变化，0 表示没有变化，T 为阈值。即计算第 k 帧和第 $k+1$ 帧中所有像素的灰度或亮度的差值，如果差值大于一个给定的值 T，那么就认为该像素发生了变化，然后通过统计图像中发生变化的像素数占总像素数的百分比来确定是否发生了镜头改变。这种方法检测镜头的边界简单快捷，但是对噪声比较敏感，容易错判含运动的场景。

2. 基于直方图的镜头分割方法

该方法的基本思路是：先把整个颜色空间(如 R,G,B)量化为 N 个槽，然后统计每个槽内含有的图像像素数，并进行规一化处理，就可以得到图像的颜色直方图，之后两幅图像的差异度就可以通过计算它们的直方图差值求得。

同一镜头内的相邻帧一般都有全局的视觉相同的元素，也就是同一镜头内相邻的帧具有相似的颜色空间分布；反之，不同镜头的相邻帧之间的颜色空间分布相似度很低。反映在直方图上就是：同一镜头内相邻帧之间的直方图差异较小；不同镜头中帧之间的直方图差异较大。显然理论上还是会存在视觉内容并不相似，直方图却相差较小的情况，但是在实际的视频序列中，出现这种情况的概率是非常小的。相对来说，基于直方图差异是比较简单也比较有效的方法，被广泛采用，但是基于直方图的方法并没有考虑图像中像素的空间信息，因此对于较缓慢的物体运动并不敏感，能减少物体运动带来的无效检测。

假设 $H_k(i)$是第 k 帧中第 i 个灰度级直方图的值，i 的范围为$[0,N]$，其中 N 为灰度级数。基于灰度直方图算法中，第 k 帧和 $k+1$ 帧的帧间差为

$$Z(k,k+1)=\sum_{i=0}^{N-1}|H_k(i)-H_{k+1}(i)| \tag{9-4}$$

当 $Z(k,k+1)$大于一个给定的值 T 时，则认为两帧图像间存在比较大的差异。

对于彩色图像通常会采用三个直方图来表示，分别是红色、绿色和蓝色直方图。但在

帧图像中，有些颜色分量所占比重比较大，在计算帧间差时此颜色分量应给予较大的权值。因此有了带权直方图：

$$Z(k,k+1)=\frac{r}{s}Z_{\text{red}}(k,k+1)+\frac{g}{s}Z_{\text{green}}(k,k+1)+\frac{b}{s}Z_{\text{blue}}(k,k+1) \qquad (9\text{-}5)$$

$$s=(r+g+b)/3$$

其中：r、g、b 分别代表图像中红色分量、绿色分量以及蓝色分量的亮度值。

直方图的方法实际上是一种基于颜色量的统计方法，因此统计结果中不再含有图像的运动、边缘、形状等信息，因此虽然这种方法对运动等不敏感，但是也就意味着许多从视觉上感觉并不太相像的图像在直方图中却可能表现得非常相似。另外一方面，虽然直方图维数越高越能反映颜色的统计信息，但是计算相似度也就越复杂，另外若需要把这些颜色直方图信息保存下来，也就越花费存储空间，如 92 维的直方图比 48 维的直方图需要多花三倍的存储空间，在存储和读取海量视频视觉特征时将不得不考虑这一点。

3. 基于块的镜头分割方法

基于块的方法是对直方图方法的一种改进。其基本思路是：将图像划分为 R 块，通过计算两幅图像中对应块的特征差值来计算它们的差异。因此基于块的第 k 帧与第 $k+1$ 帧的差值可用下式求得：

$$Z(k,k+1)=\sum_{i=1}^{R}w_i\times Z_i(k,k+1) \qquad (9\text{-}6)$$

式中 w_i 为第 i 个块上的差值权重因子。同样当 $Z(k,k+1)$大于一个给定的值 T 时，则认为两帧图像间存在比较大的差异。与根据整幅图像的特征差值来计算的直方图法相比，基于块的方法有许多优点：图像间比较的是局部特征，有利于限制噪声以及运动等带来的影响，增强了算法的鲁棒性；权重因子 w_i 可调，可以通过调节各个块的权重值，实现视频图像特定区域的特征分析和差值比较。这种算法可以在一定程度上改善对局部运动的容忍度。

4. 基于边缘改变比例的镜头分割方法

基于边缘特征方法的基本思路是：如果发生镜头变换，那么前后帧的边缘会有很大变化。在检测当前帧中的边缘是否在后一帧中消失时，只需判断在后一帧对应位置的附近是否可以找到与该边缘相匹配的边缘。在当前帧中的每个边缘经过如上检测后，后一帧中仍未得到匹配的所有边缘即被认定为新出现的边缘。不同的镜头变换对应不同的边缘描述，一般用边缘变化率(edge ch'ange ratio)来描述边缘变化特性。边缘变化率(edge change ratio，ECR)定义如下：

$$\mathrm{ECR}_k = \max\left(X_k^{\mathrm{in}}/\sigma_k, X_{k-1}^{\mathrm{out}}/\sigma_{k-1}\right) \tag{9-7}$$

其中 σ_k,σ_{k-1}分别为第 k 帧中所有边缘的像素数、第 $k-1$ 帧所有边缘的像素数。X_k^{in},X_{k-1}^{out}分别为在第 k 帧中进入边缘的像素数、第 $k-1$ 帧从图像中消失的边缘的像素数。部分研究中使用 Canny 算子进行边缘检测,为了使边缘特征能抵抗物体运动的干扰,通常在帧图像中出现的边缘像素,如果在后续相邻帧图像中一定范围内出现,则不认为该边缘像素为进入或消失的边缘像素。

9.3.2　压缩域中镜头分割方法

1. 基于 DCT 系数的镜头分割方法

DCT 系数是由 8×8 的图像块直接进行离散余弦变换得到,所以从像素域算法到压缩域算法,很容易就会想到基于 DCT 系数的方法。该方法就是利用图像块对应的 64 个系数来实现基于压缩域的镜头边界检测的,如下式所示:

$$D(f_i, f_j, k) = \frac{1}{64}\sum_{n=1}^{64}\frac{|c(f_i,k,n) - c(f_j,k,n)|}{\max(c(f_i,k,n), c(f_j,k,n))} \tag{9-8}$$

式中,$D(f_i,f_j,k)$为第 i 帧和第 j 帧的第 k 块的归一化的平均绝对差值;$c(f_i,k,n)$为第 i 帧第 k 块的第 n 个系数。

假如,$D(f_i,f_j,k)$的值大于给定的阈值 T_1,则判定为第 k 块发生了很大的变化,统计发生变化的总块数,如果也大于给定的阈值 T_2,判定在 i 帧和 j 帧之间发生了镜头变化。

2. 基于 DC 系数的方法

该方法的思路是:先构造每一帧的直流(DC)系数,即获取直流图像帧。其中,I 帧的 DC 系数直接通过帧内解码得到,而对于 B、P 帧,则可以通过 I 帧的 DC 系数和它们之间的预测信息估计出来,然后计算这些 DC 图像之间的差异度,从而检测镜头的边界。DC 图像的差值可由下式计算:

$$D_{m,n}(f_m^{\mathrm{DC}}, f_n^{\mathrm{DC}}) = \sum_{i=1}^{M} |C(f_m^{\mathrm{DC}}, i) - (f_n^{\mathrm{DC}}, i)| \tag{9-9}$$

式中,f_m^{DC} 表示第 m 帧的 DC 图像系数,M 为图像内的总块数,$C(f_m^{\mathrm{DC}},i)$表示图像 f_m^{DC} 中的第 i 个块的 DC 系数。

3. 基于运动矢量和宏块预测信息的方法

基于 MPEG 压缩域中运动矢量和宏块预测信息的方法是另一种重要的镜头边界检测方法。该方法的思路是:在一个镜头内,相机或物体的运动基本趋于稳定,因此 MPEG(动态图像专家组,一种图像压缩标准)流中的运动矢量也保留着一定的一致性,通过统计

MPEG 压缩域中的这些运动矢量信息(如预测时产生的能量差)和预测宏块信息(比如预测方向、预测数量)以检测镜头变化的边界。由于 P 帧和 B 帧的编码信息本身就代表了与预测帧之间的差异,因此只需统计这些预测信息即可,以 MPEG 压缩域中的 P 帧为例,它的预测帧间差异度可以表示为

$$D_{m,n}(f_m,f_n)=\frac{\sum_{i=1}^{N_P}E(f_m,f_n,i)}{M} \tag{9-10}$$

$E(f_m,f_n,i)$表示第 i 个预测宏块的预测能量差,该值可以通过解码运动矢量求得,当两帧图像运动矢量差别越大时,该值越大,反之越小;N 为发生的预测宏块数,M 为总宏块数。

9.4 镜头切换

当视频内容发生变化时,会出现镜头的切换。镜头切换主要有切变和渐变两种方式。

切变是指一个镜头与另一个镜头之间没有过渡,由一个镜头瞬间直接转换到另一个镜头的方法,即一个镜头猛然切换到另一个镜头,中间没有时间上的延迟,也称直接转换或突变。

渐变是指一个镜头到另一个镜头渐渐过渡的过程,没有明显的镜头跳跃。渐变包括淡入、淡出、溶解、渐变等。

(1) 淡入是指画面逐渐加强的方式。

(2) 淡出是指画面逐渐消失的方式。

(3) 溶解是指一个画面逐渐消失的同时另一个画面逐渐出现的方式,即前一帧图像里面的图片慢慢衰减,而后一帧图片缓慢变亮,直到后一帧的图片出现。

(4) 渐变是指图像从画面的某一部分开始逐渐地被另一个画面取而代之的方式,即后一帧图像的像素按照一种固定的模式替代前一个镜头的像素,如一行从右边界开始一次取代的像素点的模式。

镜头切变检测方法的基本思想是通过对比相邻图像帧之间的特征是否发生了较大变化来判断镜头的边界。由于切变镜头发生切换的相邻两个帧之间差别很大,所以无论在像素域还是在压缩域,检测突变的方法都比较成熟,检测成功率也很高。主要有基于全局特征的切变检测、基于局部特征的切变检测等。基于全局特征的切变检测将整幅图像看做一个单元计算亮度,不管是场景亮度或颜色的改变,还是目标或背景的运动,边缘轮廓的变化等都会造成亮度的突变。基于局部特征的切变检测对图像的不同部分分别对待,

最常用的方法是考虑图像中的边缘或轮廓信息。

镜头渐变的检测比切变检测复杂很多,至今仍没有取得和切变检测效果一样的成果,方法主要有阈值法、光流法和模型法等。阈值法的思路是两个镜头之间的切换是缓慢进行的,帧间差虽然有所增大,但没有一个明显的峰值,而是在一定的阈值范围之内。光流法的原理是镜头渐变切换时没有光流,而镜头运动应适合某种特定的光流类型。模型法是利用视频编辑模型来进行镜头边界检测。视频的编辑模型主要有简单色彩编辑模型、复合色彩编辑模型和空间编辑。

9.5　关键帧提取及语义提取

一个镜头包含大量信息,在视频结构化的基础上,依据镜头内容的复杂程度选择一个或多个关键帧代表镜头的主要内容,因此关键帧(或关键帧序列)便成为对镜头内容进行表示的手段。关键帧的选取一方面必须能够反映镜头中的主要事件,因而描述应尽可能准确完全;另一方面,为便于管理,数据量应尽量小,且计算不宜太复杂。

9.5.1　关键帧提取的基本原理和准则

由于在视频序列中相邻帧一般具有相似性和连续性。这样可构造出关键帧提取的基本原理:如果将所有视频帧重叠起来(在图像坐标系下),那么一个镜头中所有视频帧的特征矢量在其特征空间中形成一个轨迹。轨迹上的关键特征值所对应的帧即为关键帧。据此,关键帧提取的过程可抽象为两步:第一步,寻找图像中某特征的量化参数;第二步,判断该特征量化的参数是否为关键的特征值。

当前一般采用保守原则来选取关键帧,即关键帧的选取“宁错勿少”。在代表特征不具体的情况下,以去掉冗余帧为原则。当需要提取多幅关键帧时,关键帧提取主要是考虑它们之间的不相关性。

9.5.2　关键帧提取的方法

关键帧提取的方法主要分为两类:基于全图像序列的方法和基于压缩视频的方法。目前大多数关键帧的提取研究是基于全图像视频分析的。具体实现方法的区别主要在于检测方法的应用、特征的选择以及帧图像子块的划分。

1. 基于镜头边界的方法

该方法将镜头中的第一帧和最后一帧(或中间帧)作为关键帧。该方法简单易行,适

于内容活动性小或内容保持不变的镜头。但未考虑镜头视觉内容的复杂性,限制了镜头关键帧的个数,提取的关键帧代表性不强,效果不够稳定。

2. 基于内容分析的方法

该方法基于每一帧的颜色、纹理等视觉信息的改变来提取关键帧。比较经典的方法是帧平均法和直方图平均法。帧平均法是在镜头中计算所有帧在某个位置上像素值的平均值。然后将镜头中该点位置的像素值最接近平均值的帧作为关键帧。直方图平均法是将镜头中所有帧的统计直方图取平均,然后选取与该平均直方图最接近的帧作为关键帧。

这两种方法计算简单,所选取的帧具有平均代表意义,但选取固定数目的关键帧,无法描述有多个物体运动的镜头。于是,依据帧间内容的显著变化来选取多个关键帧的算法被提出,其基本思想是:首先把镜头的第一帧作为关键帧,然后计算前一个关键帧与剩余帧之差(用特征信息之间的距离度量),如果差值大于某一阈值,则再选取一个关键帧。这种方法可以根据镜头内容的变化程度选取相应数目的关键帧,但所选取的帧不一定具有代表意义,而且在有镜头运动时,容易选取过多的关键帧。

3. 基于光流的运动分析法

此方法是根据运动信息提取关键帧。代表算法是 Wolf 提出的运动极小值算法。Wolf 通过光流分析来计算镜头中的运动量,在运动量取局部最小值处选取关键帧。

首先用 Horn-Schunck 法计算光流。对每个像素光流分量的模求和,作为第 k 帧的运动量 $M(k)$,即

$$M(k) = \sum_{i} \sum_{j} | O_x(i,j,k) | + | O_y(i,j,k) | \tag{9-11}$$

其中,$O_x(i,j,k)$和 $O_y(i,j,k)$分别是帧 k 内像素(i,j)光流的 X、Y 分量。

然后寻找 $M(k)$的局部最小值。从 $k=0$ 开始,扫描 $M(k)-k$ 曲线,找到两个局部最大值 $M(k_1)$和 $M(k_2)$,$M(k_2)$的值与 $M(k_1)$的值至少相差 $p\%$(由经验决定),如果 $M(k_3)=\min(M(k))$,$k_1<k<k_2$,则把 k_3 选为关键帧。然后把 k_2 作为当前的 k_1,继续寻找下一个 k_2。

该法可以根据镜头的结构选择相应数目的关键帧。但其依赖于局部信息,鲁棒性不强;也没有足够重视由累加动态带来的内容变化;计算量较大。

4. 基于聚类的方法

镜头聚类是研究镜头间的关系,即如何把内容相近的镜头组合起来,需要对视频进行更高层的抽象,将内容上有关系的镜头结合起来,以描述视频节目中有语义意义的事件或活动。通过镜头聚类将镜头中的帧序列分到各个簇后,再选择视频关键帧。

基于聚类的方法是目前关键帧提取的主流技术，其基本思想是：首先确定一个初始类心，然后根据当前帧与类心的距离来判断当前帧是归为该类还是作为新的类心。将镜头中帧分类后，取各类中与类心距离最近的帧作为关键帧。

例如，设某个镜头 S_i 包含 n 个图像帧，可以表示为：$S_i=\{F_i(1),F_i(2),\cdots,F_i(i)\}$，其中 $F_i(1)$ 为首帧，$F_i(n)$ 为尾帧。根据某个图像特征（例如颜色直方图），定义两帧之间的相似度，相似度通常取为距离函数，并预先设置一个相似度阈值，以控制聚类的密度。

计算当前帧 $F_i(j)$ 与现存某个聚类质心间的距离，如果大于阈值 T，则该帧与聚类之间距离较大，不能加入该聚类。如果 $F_i(j)$ 与所有现存聚类质心间的距离均大于 T，则以 $F_i(j)$ 为质心形成一个新聚类。否则，将 $F_i(j)$ 加入到与之相似度最大的聚类中，使该帧与这个聚类的质心之间的距离最小，并且对该聚类质心做如下调整：

$$\text{centrod}' = \text{centrod} \times \frac{F_n}{F_n+1} + \frac{1}{F_n+1} \times F_i(j) \tag{9-12}$$

其中 centrod、centrod′ 和 F_n 分别是聚类原有质心、聚类更新后质心和该聚类中的帧数。

通过上面的方法将镜头 S_i 所包含的 n 个图像帧分别归类到不同聚类后，就可从每个聚类中抽取离聚类质心最近的帧作为这个聚类的代表帧，所有聚类的代表帧就构成了镜头的关键帧。

在众多的聚类算法中，K 均值聚类和模糊 C 均值聚类是两个著名的聚类算法。K 均值聚类算法的分类是清晰的，每个样本被分配到一个且只此一个聚类中；模糊 C 均值聚类算法的分类是模糊的，每个样本针对每个聚类都有一个成员函数。聚类方法能有效地表示镜头内容间的相关性，但不能有效地保存原镜头内图像帧的时间顺序和动态信息。

5. 基于压缩视频的方法

上述方法都是基于全图像序列的，即在提取关键帧之前，对视频进行解压，还原成帧图像，运算量大。基于压缩域的方法是直接从 MPEG 压缩视频流上提取关键帧，无须对视频流解压或只需部分解压，降低了计算的复杂性。目前基于压缩域的方法是直接利用压缩视频数据中的某些特征来进行分析和处理，较典型的方法有以下两类。

一是利用 MPEG 压缩视频流中 I 帧信息及其频域直流分量信息进行关键帧提取。MPEG 视频流由 I 帧、P 帧和 B 帧三种类型的帧构成，并且 MPEG 视频编码要求约每 13 帧就有一个 I 帧。由于每个镜头内必然包含 I 帧，因此可以从视频流中提取 I 帧，将原

始视频流等价为由I帧构成的视频流。再利用前面分析的方法分析相邻I帧的连续性和相似性,进行关键帧的提取。

二是利用MPEG压缩视频流中已有的离散余弦变换DCT的DC系数和运动矢量MV来提取关键帧。在MPEG视频流中,I帧采用帧内编码,主要可利用的信息是离散余弦变换DCT的DC系数;P帧采用前向预测帧间编码,主要可利用的信息是运动预测用的前向运动矢量及运动补偿用的预测残差的DCT系数;B帧采用双向预测帧间编码,运动向量有前向、后向和双向运动矢量。在提取关键帧之前,首先要检测视频的变换,在确定图像组存在镜头变换后,才进行关键帧的提取。

(1) MPEG视频流中,P帧是由前面的I帧或P帧通过前向运动补偿进行编码。当镜头变换发生在P帧时,当该P帧内没有进行运动补偿的宏块数与有运动补偿的宏块数的比值出现峰值时,则认为该P帧是一个关键帧。

(2) B帧是由其前后的参考帧通过双向运动补偿来进行编码,当镜头变换发生在B帧时,当该B帧内后向运动矢量的数目与前向运动矢量的数目的比值出现峰值时,则认为该B帧是一个关键帧。

(3) 若发生了镜头变换且P帧和B帧都不是关键帧,则推断镜头变换发生在I帧,即I帧为关键帧。渐变过程是一个连续的过程,相邻帧间变化小,没有明显的峰值,渐变中的任意一帧都可作为关键帧。

9.5.3 视频语义提取

为了高效地获取视频中包含的语义信息,常用方法是基于视频字幕的方法和基于视频中的音频信息的方法。基于视频字幕的方法是将与视频相依附的字幕中获取文本信息来获取视频语义概念。视频字幕可以分为两类:场景字幕和标注字幕。场景字幕是场景的一部分,属于原始字幕,是在录制过程中环境和物体本身的文字。尽管有些场景字幕也蕴含了语义信息,但由于场景字幕出现具有很强的偶然性并且不同的场景字幕之间的差异较大,难以寻找所有场景字幕的共同特征进行识别,因此在视频语义提取中暂时不考虑这类字幕的语义信息。而标注字幕是在视频后期制作过程中合成到视频流中的,是为解释视频内容而添加进去的。因此,一般认为标注字幕是对视频流中发生的情景的描述。为视频流提供了高度概括的语义信息。综合音频特征与可视信息进行语义分类来生成视频语义描述信息,实现视频语义提取。

9.6　视频特征提取

较常用的特征大部分建立在镜头级上，视频分割成镜头、关键帧被抽取后，就要对各个镜头进行特征提取，得到一个尽可能充分反映镜头内容的特征空间，即提取镜头的颜色、纹理以及运动甚至高级语义等各种特征，形成描述镜头的特征空间。这个特征空间将作为视频聚类和检索的依据。

视频数据的特征又分为静态特征和动态特征。静态特征的提取主要针对关键帧，可以采用图像特征提取方法，如提取颜色特征、纹理特征、形状和边缘特征等，这是基于内容的图像检索的重要内容，在第 7 章已经明确阐述。因此本章只对动态特征做详细描述。

传统获取视频运动特征的方法是运动估计（motion estimation）。运动估计是指从当前帧图像中获取运动趋势和走向的过程，是数字视频防抖动（又称为稳像原理）、视频压缩编码的核心步骤。

摄像设备与被拍摄场景之间的高速相对运动，或者摄像设备的随机抖动，都会使图像发生模糊。对连续模糊的图像序列进行运动估计和运动补偿就是电子稳像系统的核心。运动估计的目的是估计出因为摄像平台的随机抖动而带来的帧间全局运动矢量和目标运动矢量，检测出的目标运动矢量将是目标的独立运动矢量与背景的全局运动矢量的矢量之和。利用计算出的运动矢量，根据前一帧对当前帧进行运动补偿，以获得清晰稳定的图像序列。各种空间域和变换域的运动估计方法，都已经用于电子稳像中的运动估计。

完整的电子稳像系统主要是由图像预处理、运动估计和运动补偿三部分组成，运动估计又分为局部运动估计和全局运动估计。首先，由摄像机采集原始图像输入到稳像系统中，对输入的图像进行预处理（主要是平滑去噪和图像增强等处理）；其次，通过对图像当前帧与上一帧进行分析，得到对应的局部运动矢量 LMV（local motion vector），因为图像序列会包含运动主体、背景、噪声、畸变等干扰因素，需要排除这些干扰因素引起的不符合实际情况的局部运动矢量，通过余下的多个局部运动矢量估算出全局运动矢量 GMV（global motion vector）；最后通过运动平滑、运动滤波等得到运动参数，根据运动参数将图像按相反的方向移动，使得原图的抖动得以抵消，从而实现运动补偿处理，输出稳定清晰的视频图像。

图像中的物体运动可以用平移运动 $v=(v_x, v_y)$ 和旋转运动 $\omega=v_\theta$ 来表示，若不考虑摄影机的焦距变化，则第 i 帧相对于第 $i-1$ 帧的全局运动矢量 $\boldsymbol{D}$ 可以表示为 $\boldsymbol{D}_i=\{d_{x,i}, d_{y,i}, d_{\theta,i}\}$。

9.6.1 全局运动矢量的计算方法

1. 均值法

均值法是最简单的一种全局运动矢量计算方法，通过对图像中的若干个局部运动矢量进行均值计算，从而得到全局运动矢量，如下式所示：

$$\mathrm{GMV} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{LMV}_i \tag{9-13}$$

其中 LMV_i 表示区域 i 的局部运动矢量。均值法的优点是计算简单、速度快，当图像中没有运动主体的时候效果接近最优解，但是其中每个 LMV_i 具有相同的权重值，在图像中出现干扰现象的时候无法剔除这些干扰，尤其是图像中出现快速运动的小目标主体时，全局运动矢量将受到很大的影响。

2. 权重法

在均值法中所有局部运动矢量具有相同的权重，这与稳像效果的目标不符，数字稳像要求在视野范围内尽量多的物体保持稳定，即 MPC 准则。因此根据上述准则，提出了一种对 LMV 赋予权重值的方法，其中权重是从该 LMV 的稳定度和隔离度两方面来衡量的。

稳定度用于描述 LMV 在前后帧之间的关系，当两帧之间的 LMV 具有紧密关联时，对其赋予较高的权重值，否则赋予一个较低的权重值。

隔离度用于描述 LMV 与所有 LMV 均值之间的关系，当某个 LMV 值与均值相差较大时，认为其受到运动主体或是其他因素的影响，并给其赋一个较低的权重值，这样将降低该 LMV 对 GMV 的影响。

权重法的实现如下。

定义块 $B_i(i=1,\cdots,M)$ 的局部运动矢量为 $\mathrm{LMV}(i)=(x_1,y_1)$，全局运动矢量为 $\mathrm{GMV}=(x,y)$，贝隔离度 I_i 可由下式表示：

$$I_i = |x_i - m_x| + |y_i - m_y|, \quad i = 1,\cdots,M \tag{9-14}$$

其中，

$$m_x = \frac{\sum x_i}{M}, \quad m_y = \frac{\sum x_i}{M}$$

稳定度 S_i，可由下式得出：

$$S_i = |x_i - x_{\mathrm{old}}| + |y_i - y_{\mathrm{old}}| \tag{9-15}$$

其中$(x_{\mathrm{old}},y_{\mathrm{old}})$表示上一帧的全局运动矢量。

3. 运动估计数学模型

视频图像序列的抖动是由摄像机的随机抖动造成的,因此需要分析摄像机系统运动的类型、摄像机运动与图像运动的关系。对于不同的视频图像序列帧间的运动采用不同的变换模型,常用的三种模型有:平移模型、相似模型和反射模型。

(1) 平移模型。只分析图像的平移运动,此模型可表示为

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} \mathrm{d}x \\ \mathrm{d}y \end{bmatrix} \tag{9-16}$$

式中,(x', y')和(x, y)分别表示图像序列当前帧和参考帧中对应像素点的坐标值,$(\mathrm{d}x, \mathrm{d}y)$为当前帧相对于参考帧在 x 和 y 方向上的位移量。

(2) 相似模型。考虑图像的旋转和变焦两种运动时,此模型可表示为

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = s \times \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} \mathrm{d}x \\ \mathrm{d}y \end{bmatrix} \tag{9-17}$$

式中,(x', y')和(x, y)分别表示图像序列当前帧和参考帧中对应像素点的坐标值,$(\mathrm{d}x, \mathrm{d}y)$为当前帧相对于参考帧在 x 和 y 方向上的位移量,s 和 θ 分别表示当前帧相对于参考帧的缩放系数和旋转角度。

(3) 反射模型。考虑图像出现扭转变化情景时,上述两种模型不能够反映变换的情况,此时模型可表示为

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = s \times \begin{bmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} \mathrm{d}x \\ \mathrm{d}y \end{bmatrix} \tag{9-18}$$

式中 $k_{11}, k_{12}, k_{21}, k_{22}$分别为旋转参数,$s$ 为变焦系数,(x', y')和(x, y)分别表示当前帧和参考帧中对应像素点的坐标,$(\mathrm{d}x, \mathrm{d}y)$分别表示当前帧相对于参考帧在 x 和 y 方向上的位移量。

如果要用数学模型更加精确地描述摄像机的运动,需要更多的模型参数,相应的计算复杂性越高。计算复杂性越高,实时性越差,应综合考虑精确度与实时性,根据不同的情况,选择合适的数学模型实现数字稳像。

9.6.2 视频运动估计

视频运动估计的基本思想是将图像序列的每一帧分成许多互不重叠的宏块,并认为宏块内所有像素的位移量都相同,然后对每个宏块到参考帧某一给定特定搜索范围,根据一定的匹配准则找出与当前块最相似的块,即匹配块,匹配块与当前块的相对位移即为运动矢量。视频压缩的时候,只需保存运动矢量和残差数据就可以完全恢复出当前块。本节只说明视频图像运动矢量估计的两种较简单的情况:平移运动估计和旋转运动估计。

1. 平移运动估计

平移运动是相邻帧间特定的像素的移动，可通过光流法、块匹配法、特征匹配法等其他方法得到。

图 9-4 是块匹配法的平移运动估计示意图。其中蓝色虚线区域表示搜索范围，黑框区域表示分割的块图像，红色框区域表示上一帧中黑框区域的图像，黑框与红框的中心位移矢量即为局部运动矢量。

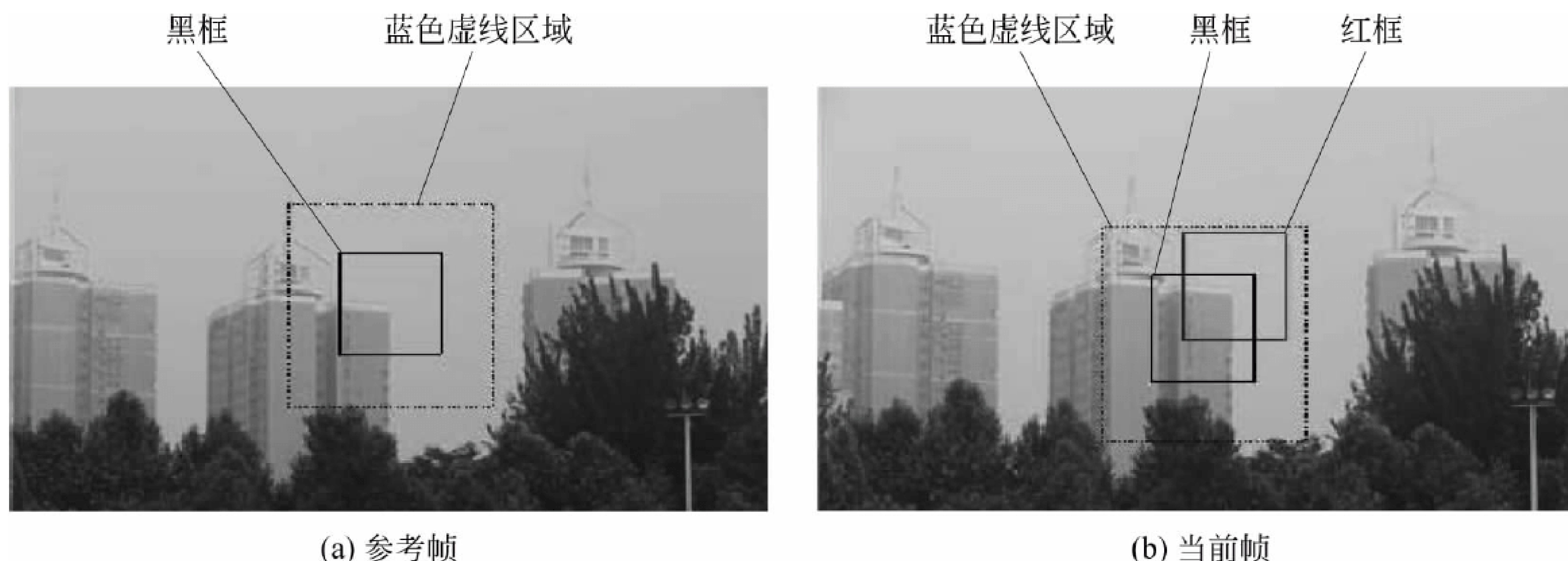


图 9-4　视频图像运动估计搜索示意图

2. 旋转运动估计

从平移运动矢量中，可通过运动模型得到旋转矢量。当连续图像序列中的点(x^1, y^1)以(x_0, y_0)为圆心旋转时，只讨论纯粹的旋转运动，则我们表示转动之后点的坐标(x^2, y^2)为

$$\begin{bmatrix} x^2 \\ y^2 \end{bmatrix} = s \times \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x^1 \\ y^1 \end{bmatrix} + \begin{bmatrix} \mathrm{d}x \\ \mathrm{d}y \end{bmatrix} \tag{9-19}$$

其中 s 为缩放系数，在摄像机大致固定、焦距不变的系统中，可以将 s 看做 1，则可得

$$\begin{bmatrix} x^2 \\ y^2 \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x^1 \\ y^1 \end{bmatrix} + \begin{bmatrix} \mathrm{d}x \\ \mathrm{d}y \end{bmatrix} \tag{9-20}$$

当旋转角 θ 较小时，则可表示为

$$\begin{bmatrix} x^2 \\ y^2 \end{bmatrix} = \begin{bmatrix} 1 & -\theta \\ \theta & 1 \end{bmatrix} \begin{bmatrix} x^1 \\ y^1 \end{bmatrix} + \begin{bmatrix} \mathrm{d}x \\ \mathrm{d}y \end{bmatrix} \tag{9-21}$$

对于上式有 N 个匹配的点对，这样就获得了 $2N$ 个线性方程组成的三个未知数 θ、Δx 和 Δy 的方程组。以矩阵 $\boldsymbol{b}=\boldsymbol{Ax}$ 的形式重新排列，可得

$$\boldsymbol{b}=\begin{bmatrix} x_{i1}-sx_{j1} \\ y_{i1}-sy_{j1} \\ \vdots \quad \cdots \\ x_{iN}-sx_{jN} \\ y_{iN}-sy_{jN} \end{bmatrix};\quad \boldsymbol{A}=\begin{bmatrix} -sy_{j1} & 1 & 0 \\ +sx_{j1} & 0 & 1 \\ \vdots & \cdots & \\ -sy_{jN} & 1 & 0 \\ +sx_{jN} & 0 & 1 \end{bmatrix};\quad \boldsymbol{x}=\begin{bmatrix} \theta \\ \Delta x \\ \Delta y \end{bmatrix} \Rightarrow \boldsymbol{x}=(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\boldsymbol{b} \tag{9-22}$$

尺度因子 S 可以通过拟合相似模型单独计算出来，然后代入上式估计剩余参数。对于平移和旋转变换 s 是常量，反比于时刻 t_i 和 t_j 的任意两帧图像的距离。因此，可以通过计算采集于 k_i 和 k_j 的两帧图像中的匹配块集来估计 s。首先获得每个匹配块集的质心：

$$x_f=\frac{1}{N}\sum_{k=1}^{N}x_{jk},\quad y_f=\frac{1}{N}\sum_{k=1}^{N}y_{jk} \tag{9-23}$$

其中(x_f,y_f)为匹配块集 S_f 的质心坐标，(x_{jk},y_{jk})为匹配块集中块 k 的坐标。用 λ_{jk} 表示从匹配块 k 到帧 f 的质心距离，故帧 i 和帧 j 间的尺度变换因子可由下式得出：

$$\begin{bmatrix} \lambda_{i1} \\ \lambda_{i2} \\ \cdots \\ \lambda_{iN} \end{bmatrix}=s\begin{bmatrix} \lambda_{j1} \\ \lambda_{j2} \\ \cdots \\ \lambda_{jN} \end{bmatrix} \Rightarrow s=\left[\begin{bmatrix} \lambda_{j1} \\ \lambda_{j2} \\ \cdots \\ \lambda_{jN} \end{bmatrix}^{\mathrm{T}}\begin{bmatrix} \lambda_{j1} \\ \lambda_{j2} \\ \cdots \\ \lambda_{jN} \end{bmatrix}\right]^{-1}\begin{bmatrix} \lambda_{j1} \\ \lambda_{j2} \\ \cdots \\ \lambda_{jN} \end{bmatrix}^{T}\begin{bmatrix} \lambda_{i1} \\ \lambda_{i2} \\ \cdots \\ \lambda_{iN} \end{bmatrix} \tag{9-24}$$

$$\Rightarrow s=\frac{\sum_{k=1}^{N}\lambda_{jk}\cdot\lambda_{ik}}{\sum_{k=1}^{N}\lambda_{jk}\cdot\lambda_{jk}}$$

平移和旋转参数在估算尺度变换因子后，可进行计算，对于反射变换 $2N$ 个方程可写成如下形式：

$$\begin{bmatrix} r_{11} \\ r_{12} \\ \Delta x \\ r_{21} \\ r_{22} \\ \Delta y \end{bmatrix}=\left[\begin{bmatrix} x_{j1} & y_{j1} & 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{jN} & y_{jN} & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & x_{j1} & y_{j1} & 1 \\ \vdots & & & & & \\ 0 & 0 & 0 & x_{jN} & y_{jN} & 1 \end{bmatrix}\right]\begin{bmatrix} x_{i1} \\ \vdots \\ x_{iN} \\ y_{i1} \\ \vdots \\ y_{iN} \end{bmatrix} \tag{9-25}$$

由上式我们可以求出旋转和平移参数。

通过平移运动估计，我们可以得到图像中点(x,y)的运动矢量，设 $u=x_2-x_1$，

$v=y_2-y_1$，点(x^2,y^2)是(x^1,y^1)经过旋转之后的坐标。假设在纯旋转过程中，任意点的运动方向都与旋转中心的同心圆相切，那么，假设有一组点，则点的运动矢量的中垂线均相交于旋转中心。

运动矢量直线段的中垂线可以表示为 $y=k(x-x_1)+y_1$，k 为直线斜率，$k=(y_2-y_1)/(x_2-x_1)$。理论上只要两点便能确定旋转中心，但是为了保证在视频图像中出现干扰时程序的鲁棒性，通常使用多组数据进行互相匹配以获得最佳运动参数。对于 N 点的情况，有 N 组等式：

$$\begin{bmatrix} -a_1 & -a_2 & \cdots & -a_N \\ 1 & 1 & \cdots & 1 \end{bmatrix}\begin{bmatrix} x_0 \\ y_0 \end{bmatrix}=\begin{bmatrix} b_1 & b_2 & \cdots & b_N \end{bmatrix}^{\mathrm{T}} \tag{9-26}$$

则旋转中心为 $\boldsymbol{x}=(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{b}$。找到旋转中心之后，旋转角 θ 可由下列公式得到：

$$\theta=\tan^{-1}\frac{(y^2-y_0)(x^1-x_0)-(c^2-x_0)(y^1-y_0)}{(x^2-x_0)(x^1-x_0)-(y^2-y_0)(y^1-y_0)} \tag{9-27}$$

若考虑图像缩放系数，可得

$$s\times\sin\theta=\frac{y_1-y_2\dfrac{y'_1-y'_2}{x'_1-x'_2}(x_1-x_2)}{x'_1-x'_2+\dfrac{(y'_1-y'_2)^2}{x'_1-x'_2}} \tag{9-28}$$

则图像指定像素点位移可表示为

$$\begin{bmatrix} \mathrm{d}x \\ \mathrm{d}y \end{bmatrix}=\begin{bmatrix} x_i \\ y_i \end{bmatrix}-s\times\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}\begin{bmatrix} x'_i \\ y'_i \end{bmatrix} \tag{9-29}$$

9.6.3 运动矢量估计的常用算法

运动矢量估计的研究总是围绕着解决计算复杂度和检测精度这对矛盾进行。运动矢量估计的算法主要有灰度投影法、特征匹配法、光流法、块匹配法等，应该根据实际需要合理选择运动估计算法。

1. 块匹配法

1）块匹配运动估计原理

视频图像序列的相邻帧间存在很大的时间冗余，对视频序列进行压缩时，采用各种运动估计算法，可以大幅度提高视频编码的效率。块匹配法因其简单有效，在视频编码中得到广泛应用。块匹配运动估计法是基于块内各像素运动一致性的假设基础上的。

块匹配法的原理为：将图像的当前帧划分为固定大小（$M\times N$ 像素）的图像子块，一般

是 16×16 或者 8×8 像素，并假定位于同一图像子块内的所有像素具有相同的位移，然后对当前帧中的每一块，在上一帧的一定范围内（搜索窗口），根据一定的匹配准则找出最优匹配块（预测块），该块就是从上一帧最优匹配块位置处平移过来的，所得运动位移即为当前块的运动矢量。设可能的最大位移矢量为（dx，dy），则搜索范围为（M+2dx）×（N+2dy），为了方便算法的实现，子块的 M 和 N 取值一般相等，dx 和 dy 也取相等。预测块和当前块的像素差值组成残差块，预测块与当前块之间通过匹配准则函数得到的值称为块匹配误差。

块运动模型分为块平移模型和可变形块模型两种，块平移模型假定每个块只做二维平移运动。给定两帧视频图像 $\psi_1(x)$和 $\psi_2(x)$，对于 $\psi_1(x)$中的一个块 β，可由 $\psi_2(x)$中一个同样大小的块重建，即

$$\psi_1(x)\mid_{x\in\beta}=\psi_2(x+d) \tag{9-30}$$

其中，d 是两个块的空间距离。块重建的过程称为运动补偿。式（9-30）中的块可以是重叠或者非重叠的。对于非重叠块，每个块估计一个运动矢量，由式（9-30）进行运动补偿；对于重叠块，重叠部分像素的运动矢量可以由两个块的平移矢量平均得到，也可选择其中匹配程度较好的运动矢量。

可变形块运动模型能够实现对物体的旋转、缩放、变形等建模。块的运动参数不再是简单的一个平移参数，而是空间变换参数，常用的可变形块运动模型有投影运动、仿射运动、双线性运动等。使用可变形块能更准确地找到匹配位置，并且对于锐体和旋转物体的匹配具有较好的效果，但是使用可变形块将占用大量系统资源，并且在实际匹配过程中效果提升并不明显，为了计算和分割方便，在实时稳像系统中通常使用固定形状的块，一般为正方形。

2）块匹配运动估计的技术指标

块匹配运动估计的效率主要体现在图像质量、压缩码率、搜索速度三方面。运动估计越准确，预测补偿的图像质量就越高，补偿的残差就越小，补偿编码所需位数越少，比特率越小；运动估计速度越快，越有利于实时应用。提高图像质量，加快估计速度，减少比特率是块匹配运动估计的目标。块运动估计可以从以下四个方面进行：块形状与大小、块匹配准则、初始搜索点的选择、算法的评价指标。

（1）块形状与大小。块匹配方法隐含着如下假设：同一块内像素的运动是一致的。显然该假设具有一定的片面性，但选择合适的块形状与大小可在一定程度上消除这种片面性。一般来说，块形状选用正方形是比较自然的选择，这样既便于图像的划分，又有利

于块匹配准则函数的计算。但这并不一定是最佳选择,有的算法采用了其他形状,如三角形等。块大小的选择受两个矛盾的约束:块大时,块内各个像素做相等平移运动的假设不合理;快越小,编码一帧图像所需要的运动估计次数越多,因而需要存储和传输的运动矢量数也越多,则编码效率降低。因此,要综合考虑多种因素,选择合适的块大小。作为折中,通常选择 16×16 的宏块作为单位。

(2) 块匹配准则。块匹配准则是判断块相似程度的依据,因此匹配准则的好坏直接影响了运动估计的精度;另一方面,匹配运算复杂度、数据读取复杂度在很大程度上取决于所采用的块匹配准则。因此,提高运动估计算法的速度可以用两种途径:一种是减少搜索匹配的点数,另外一种是降低块匹配准则的计算复杂度。运动估计算法中常用的匹配准则有以下两种

① 平均绝对误差(mean absolute difference criterion,MAD)

$$\mathrm{MAD}(\vec{d}) = \frac{1}{N_1 N_2} \sum_{(n_1, n_2) \in B} | s(\vec{n}, k) - s(\vec{n} + \vec{d}, k+1) | \tag{9-31}$$

MAD 准则实现简单、方便,所以使用最多,还可以将 MAD 简化为 SAD(sum of absolute difference),即绝对误差求和,可以去掉不必要的运算。SAD 定义为

$$\mathrm{SAD}(\vec{d}) = \sum_{(n_1, n_2) \in B} | s(\vec{n}, k) - s(\vec{n} + \vec{d}, k+1) | \tag{9-32}$$

$$\mathrm{SAD}(\vec{d}) = M \times N \times \mathrm{MAD}(\vec{d}) \tag{9-33}$$

② 均方误差(mean square error,MSE)

$$\mathrm{MSE}(\vec{d}) = \frac{1}{N_1 N_2} \sum_{(n_1, n_2) \in B} [s(\vec{n}, k) - s(\vec{n} + \vec{d}, k+1)]^2 \tag{9-34}$$

$$\vec{d} = (d_1, d_2), \quad \vec{n} + \vec{d} = (n_1 + d_1, n_2 + d_2)$$

(3) 初始搜索点的选择。一种是直接选择参考帧对应的(0,0)位置,这种方法简单,但容易陷入局部最优。如果采用的算法初始步长太大,而原点又不是最优点,有可能使快速搜索跳出离原点周围可能性比较大的区域而去搜索远距离的点,导致搜索方向的不确定性,故有可能陷入局部最优。另一种是选择预测的起点。由于运动物体的整体相关性和视频运动的连续性,因此视频序列图像的运动必然具有时间和空间上的相关性。许多算法都利用这种相关性先对初始搜索点进行预测,以预测点作为搜索起点。大量的实验证明,预测点更加靠近最优匹配点,即加强了运动矢量中心偏置分布,使得搜索次数减少。

(4) 算法的评价指标。运动估计算法的优劣,主要取决于匹配效果和搜索时间。匹

配效果可以通过人眼进行主观评价，但这具有较大的随意性，且不易进行定量的比较。一般选择平均峰值性噪比(PSNR)或者平均 MSE 进行评价。

① 信噪比：

$$\mathrm{PSNR} = 10\log_{10}(255^2/\mathrm{MSE}) \tag{9-35}$$

② 搜索时间：由于搜索时间受运动平台及其他因素的影响，目前常见的还是比较搜索点数即搜索过程中进行匹配的次数。对于块匹配运动估计，计算复杂度主要依赖于平均搜索点数。

2. 灰度投影法

视频图像序列的实质是灰度发生连续变化的一组图像，灰度投影法就是利用图像的灰度分布变化特性获得图像的全局运动位移矢量，这与块匹配法利用单像素信息先获得小块的局部运动矢量后获得全局运动矢量不同。灰度投影法是利用图像序列的行列各自投影曲线做互相关处理，进而获得图像序列的全局运动矢量。因此和块匹配法相比，灰度投影法计算量少。

除了具有以上优点外，灰度投影法还有一些缺点：①当视频图像序列的灰度对比不明显时，则不易实现投影曲线的互相关运算，得出的全局运动矢量精度不高。②场景中物体的局部运动对投影曲线的互相关运算会影响投影法的精度，也会使得全局运动矢量精度降低。③用全相关搜索相关曲线的峰值时会产生较多的时间浪费。若直接使用投影算法对整幅图像进行运动估计，会得出不准确的全局运动矢量，从而影响稳像系统的性能，因此灰度投影法一般要进行预处理。灰度投影法主要包括图像映射、投影滤波、相关计算三个步骤。

(1) 图像映射：把每一帧输入的初始二维图像映射为两个独立的一维波形。

(2) 投影滤波：当帧间运动量大时，边缘信息在每一帧图像上是唯一的，因此边缘信息在互相关计算时会对互相关的峰值产生不利影响。为解决此问题，需通过余弦滤波器进行滤波处理，降低边缘信息的幅值，保留中间区域的幅值。

(3) 相关计算：将得到的投影图与参考图像的投影图做相关计算，在相关曲线中的唯一峰值即为运动矢量所求的位移值。

3. 特征匹配法

特征匹配的基本原理是：通过在参考帧中选取典型特征作为标识，并在当前帧中以一定的匹配准则进行搜索，以寻找对应的特征结构，从而获得图像序列的全局运动矢量。特征匹配法的步骤如下。

(1) 从参考帧中提取特征量。通常特征量应该具有这些特点：有比较高的定位精

度、在图像中尽可能分布均匀、有比较丰富的图像信息、与周围特征比较有一定的独特性。

(2) 进行特征匹配。按照一定的匹配准则,在当前帧中进行特征量匹配。

(3) 剔除伪匹配特征量。由于图像序列中存在的物体移动、遮挡等因素,会出现找不到特征量的情况,此时会出现伪匹配特征量,为了防止全局运动矢量的降低必须剔除这些伪特征量。

(4) 全局运动矢量的确定。把获得的局部运动矢量代入对应的数学模型得到全局运动矢量。

在特征匹配过程中,应该选取明显的局部特征,常用的特征有角点、边缘、直线等。边缘特征匹配是常用的方法。边缘特征匹配的运动估计算法主要有两个步骤。

(1) 图像的边缘检测。通过各种边缘检测算法分别提取出参考帧图像和当前帧图像的边缘。

(2) 图像的边缘匹配。将参考帧的二值化边缘图像分为四块,在每块中选取固定数量的像素称之为核,用这个核在当前帧的边缘图像中搜索对应的区域,根据最小绝对误差MAE 搜索准则来确定最佳的匹配块。对各个子块得出的局部运动矢量进行分析,采用均值滤波的方法得出全局运动矢量。

4. 光流法

观察动态物体时在视网膜上产生连续的光强度变化,就像是光的"流动"。光流是空间运动物体在观测成像面上的像素运动速率分布,反映了在一定时间间隔内由运动所造成的图像变化。光流中既包括了被观察物体的动态行为信息,也包括了有关的结构信息。它利用图像序列的像素强度数据的时域变化和相关性来确定各自像素的位置"运动",即反映图像灰度在时间上的变化与视频中物体结构及其运动的关系。通常光流由相机运动、场景目标运动或者两者的共同运动产生。每个像素都有一个运动矢量,因此可以较为准确地反映相邻帧间的运动。

光流的计算方法大致可分为三类:基于匹配的、频域的和梯度的方法。

(1) 基于匹配的光流法包括基于特征和基于区域两种。基于特征的方法是通过不断地对目标主要特征进行定位与跟踪,此方法对大目标的运动和亮度变化具有鲁棒性,但是光流稀疏而且较难精确匹配。基于区域的方法对相似区域进行定位,通过这些区域的位移计算光流,但此方法计算的光流仍然比较稀疏。

(2) 基于梯度的方法是利用图像序列中像素强度的时域变化和相关性对图像的运动场进行估计,将相似的运动矢量合并成运动目标。根据运动目标随时间变化的光流特性,

利用图像相邻帧的差分进行图像分割，利用图像分割信息可以得到基于光流法的运动目标检测。

光流法对于图像的边缘梯度值有很高的要求，并且需要对全帧图像进行处理，当帧率较高、图像尺寸较大时，则要求计算机具有较高的计算速度。光流法运动估计的优点是无须知道当前场景信息就可以用于检测运动目标。但是，基于光流的方法利用了灰度的变化信息，光流的连续性在很大程度上依赖于光照条件和物体的反射特性。相对于块匹配法，光流法可以更为准确地反映对象的运动，分割精度高，但是计算量大，难以满足实时性检测。

9.7　视频聚类

视频聚类是研究视频流中镜头之间的关系，也就是把内容相近的镜头重新组合在一起，用以描述视频中有意义的事件，或是为了缩小检索的范围，提高检索的效率。

聚类算法的基本思想是使用分裂法对给定的 n 个样本、元素或记录的数据集，使用分裂法构造 k 个分组，每个分组代表一个簇，同时 $k<n$。K 个分组满足如下条件。

（1）每个分组至少包含一个样本。

（2）每个数据样本属于且仅属于一个分组（该条件对某些模糊聚类算法可以放宽）。

对于给定的 k，算法首先给出一个初始的分组方法，以后通过反复迭代的方法改变分组，使得每一次改进之后的分组方案都比前一次好，而“好”的标准是分到同一分组的样本越接近越好，不同分组中的样本越远越好。经典的聚类方法主要有 KM、FCM 及 KHM 聚类算法。下面简单介绍 KM 聚类算法。

KM 算法的基本思想是将 n 个数据对象划分到 k 个簇，使获得的簇满足在同一簇中的对象相似度较高，而在不同簇中的对象相似度较小。聚类相似度通过利用各簇中对象的均值获得一个“中心对象”（也称质心）进行计算。

KM 算法的工作流程：首先，从 n 个数据对象任意选择 k 个对象作为初始聚类中心，对于剩下的其他对象，则根据它们与这些聚类中心的相似度或距离，分别将它们分配给与其最相似的（聚类中心所代表的）聚类；接着，计算每个新聚类的聚类中心（该聚类中所有对象的均值）；不断重复这一过程直到标准测度函数开始收敛为止。一般采用的标准测度函数是均方误差和函数，如下式：

$$\mathrm{SSE} = \sum_{j=1}^{k} \sum_{x_i \in c_j} \| x_i - u_j \|^2 \tag{9-36}$$

其中，$u_j=\frac{1}{n_j}\sum_{x_i\in c_j}x_i$ 代表簇 c_j 的均值，n_j 代表簇 c_j 的样本数。K 个簇具有以下特点：簇本身尽可能地紧凑，簇之间尽可能地分开。

KM 算法描述如下。

输入：簇个数 k，以及包含 n 个数据对象的数据集。

输出：满足均方差和最小标准的 k 个簇。

(1) 从 n 个数据对象任意选择 k 个对象作为初始聚类中心。

(2) 循环(3)到(4)，直到每个簇不再发生变化为止。

(3) 根据每个簇对象的均值(中心对象)，利用下面欧几里德距离公式，计算每个对象与这些中心对象的距离，并根据最小距离重新对相应对象进行划分。

(4) 重新计算每个簇的均值(中心对象)。

欧几里德距离(Euclidean distance)是度量两个对象之间距离的最常用的方法之一，如下式：

$$\text{Euclidean}(O_i,O_j)=\sqrt{\sum_{d=1}^{p}(o_{id}-o_{jd})^2} \tag{9-37}$$

O_i 为 p 维空间中的对象，用数值矢量 $O_i=\{o_{ij}\mid 1<j<p\}$ 表示。其中，o_{ij} 表示第 i 个数据对象的第 j 个特征的值，p 表示特征的数目。

9.8 视频结构索引

在视频数据的浏览和检索过程中，需要对视频数据进行大量随机的浏览检索、视频帧抽取、剪辑以及播放等操作，而 MPEG 码流对随机读取的支持并不好，这主要是因为以下几点。

(1) MPEG 采用差分预测编码，因此每帧编码数据的大小不固定，即使是固定比特率的 MPEG 编码流，帧之间的大小差异也非常显著。

(2) MPEG 帧的解码可能依赖于码流中的其他帧，例如 P 帧的解码依赖于在前面的 I 帧或 P 帧，B 帧的解码依赖于前面或后面的 P 帧，其实无论是 P 帧还是 B 帧的解码都依赖于它们前面第一个 I 帧的解码，否则都无法恢复出图像。

MPEG 编码数据本身并没有提供随机定位视频帧的机制，而在视频数据的浏览和检索过程中，需要大量地随机操作 MPEG 数据，这些操作又几乎都离不开视频结构索引的信息。这使得解决 MPEG 码流中的精确随机读取、建立视频结构索引的问题显得十分

突出。

9.8.1　视频结构索引的机制

在视频数据中,人们能访问到的最小单位就是图像帧,无论是视频图像或关键帧图像的随机浏览,还是随机播放或视频剪辑,都是从某一帧开始。而视频结构索引所要达到的目标就是能按需求随机定位到视频的某一帧。因此,只要能为视频数据里每一帧图像建立好索引信息,就可以在任何时候从该帧访问视频数据。

在建立 MPEG 数据流的编码结构模型和帧序列的结构模型时,得到的信息非常多,例如,视频帧在数据中的字节位置(position)可以挖掘时间信息(time-stamp)、帧序号(frame-ID)、帧类型(frame)以及帧的预测长度等。

帧序号有两种:一种是在编码数据里的编码序号,也称绝对帧号,对于任意一段视频数据,其第一帧图像的绝对帧号一般不为 1;另一种是相对帧号,即在解码过程中,可以用一个计数器累计解码的帧数,这种情况下对于任意一段视频其帧号都是从 1 开始累积的。

预测长度指当前帧与第一个 I 帧之间间隔的帧数,而在解码的时候可以统计得到每一帧在编码数据里的字节数,因此也就可以以此作为预测的字节长度。由于所有的 P 帧和 B 帧的解码都依赖于 I 帧解码,因此 I 帧的索引信息是整个视频结构索引中最基本的信息。

从上面的分析可以看出:在解码 MPEG 数据的过程中,通过建立数据流的结构模型,我们就可以建立起视频帧的索引,之后也就可以通过这些索引信息随机、快速地从任意位置开始访问视频数据。结合视频帧的索引信息和视频结构化分析的结果(如镜头起始帧号、结束帧号或字节位置,关键帧帧号或字节位置等),就可以建立起视频的结构索引,这就是视频结构索引的建立过程。

9.8.2　索引信息的存储

1. 基于文件存储的方法

在建立视频结构索引的过程中,还需考虑索引信息的存储问题。传统的索引信息存储方式是:先分析整个 MPEG 码流,把提取得到的帧索引信息临时存放于内存之中,等分析完全部 MPEG 码流以后,再将所有的索引信息存储为一个索引文件。之后就可以基于该索引文件保留的视频结构信息随机地精确访问 MPEG 数据了。

基于文件的视频结构索引虽然解决了视频数据的随机访问问题,但是也存在不足,主要体现在以下几个方面。

（1）不能在视频处理过程中浏览和检索视频处理结果，因为在视频处理完毕以前，帧的索引信息均在内存中，还没有生成索引文件，浏览终端访问不到这些信息，也就无法浏览处理结果，同样也无法检索视频的内容分析结果。

（2）基于文件索引的视频处理系统鲁棒性不够高。在实际的视频检索系统中，有可能由于某种原因（如数据的误码率比较大）导致系统没有处理完视频数据就不能再运行了，一方面对于依赖于索引文件的浏览和检索将无法进行，另一方面可能会影响到后面的视频数据的处理。

（3）索引文件的管理和维护会随着视频数据量的增加变得越来越困难。当处理大量的视频数据时，也就意味着会存在大量的索引文件，那么对这些文件的管理和维护也就会越复杂。

2. 基于数据库的方法

从上述的分析可以看到，基于文件的视频结构索引，最大的不足就是不能在视频处理的时候浏览和检索处理的结果。这对于基于文件的视频处理固然不会有太大影响，但是对于有实时浏览和检索需求的视频流处理就是不可容忍的事了，例如处理的视频流若为两个小时的新闻，那么就意味着需要两个小时以后才能浏览和检索这段新闻视频的处理结果。经大量的实际应用和研究发现，基于数据库的视频结构信息存储方法可以解决视频检索系统中的实时浏览检索问题。

该方法的出发点有两点：一是在视频检索系统中，通常是以镜头（关键帧）为基本单位进行浏览和检索，而非基于视频帧，因此无须存储所有视频帧的索引信息，只需镜头边界和关键帧的索引信息即可；二是由于P、B帧的解码均依赖于前面的I帧，通常预测长度最长为一个图组长度，这不会影响到实际应用，因此，可以把镜头边界都定位在I帧，关键帧也为I帧（镜头中的第一个或最后一个或中间一个I帧皆可）。结合以上两点，把视频结构信息存于数据库的方法，即把镜头边界和关键帧（皆基于I帧）的索引信息存于数据库中，这样做有以下优点。

（1）可实时地取出数据库中的索引信息用于浏览和检索，以满足视频处理和浏览检索的实时性要求。即只要有处理结果，就可对其浏览检索。

（2）由于只存储镜头边界和关键帧的索引信息，相当于一个镜头内仅需存几个视频帧的索引信息，因此大大减少了存储空间。

（3）可以基于数据库方便地统一管理视频信息，解决了海量视频数据的管理和维护问题。

（4）后续的视频处理不会影响到视频数据的浏览和检索，系统的鲁棒性大大提高。

9.9 视频摘要

由于视频索引是由镜头中的关键帧构成的,是静止的图像,用户有时并不能通过这些不连贯的图像得到自己想要的信息,这就需要我们对视频进行分析。

视频分析和处理的初期主要集中在分析视频帧的低层特征上,例如颜色、形状、纹理等;而目前的研究则主要集中在更加接近直观内容的分析上,其中一个重要的研究内容就是如何从原始视频中提取视频片段,同时保留比较完整的视频内容以及如何实现对视频的快速浏览和检索,这就是目前数字视频技术的一个研究热点和难点问题即视频摘要(video abstraction)。

一篇文章的摘要,就是对文章的简要总结,而视频摘要的概念则是从文本摘要延续而来的,顾名思义,视频摘要就是对一个较长的视频文件的内容所进行的一个简短的小结。视频摘要是静止图像或者是运动图像的序列(这些图像序列可以附带音频也可以不带),这个序列比原始视频要短很多,但是这个序列应保留原始视频的基本内容,以便能够实现对原始视频进行快速浏览和检索。

1. 视频摘要的分类

视频摘要就是通过对视频进行分析处理后,自动生成紧凑的能够充分表现视频语义内容的静止或者运动的图像序列。视频摘要还可根据是静止图像序列还是运动图像序列划分,可分为视频概要(video summary)和缩略视频(video skimming)两大类,其进一步细分如图9-5所示。

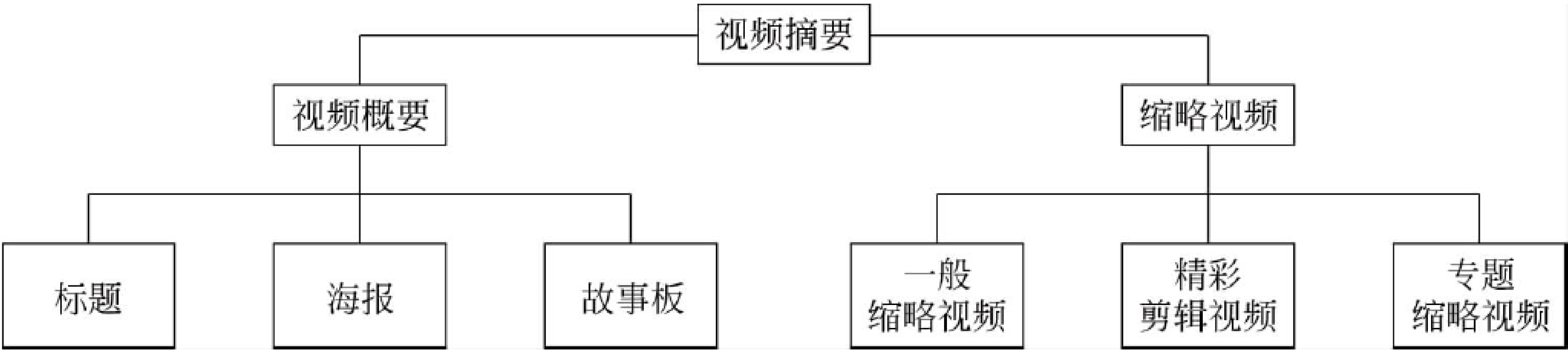


图9-5 视频摘要

对于视频概要,可以分为标题、海报和故事板三类,其中标题是对视频内容的一段简短的文字描述,这种方式虽是最紧凑最简单的视频摘要形式,但是却很难由计算机自动生成能准确概括视频内容的文字描述;海报又称为视频代表帧,它是一幅对原始视频有代表意义的图像,它可以提供直观的可视信息,但是无法表现视频的动态特性;故事板是从原

始视频中提取的,按照一定顺序和一定形式排列的多帧代表帧图像序列,它可以给用户提供视频的总体描述,在浏览中也可以方便地定位到用户感兴趣的部分。在视频概要的生成过程中,一般不需要伴音和文本的辅助,由于不存在时间同步的问题,因此不仅实现速度快,显示速度也很快。视频概要还可以用全景图拼接法来表现更加全面和准确的信息,也可以通过一定的空间关系来显示时序图像。

对于缩略视频,可以分为精彩剪辑视频、专题缩略视频和一般缩略视频三类,其中精彩剪辑视频就是前面提到的在电影电视中应用广泛的视频摘要,为了吸引观众,剪辑视频一般由原始视频中的精彩画面组成,并且不包含故事的结局;专题缩略视频是特定领域视频的摘要,专题缩略视频的实现一般都要结合该领域的知识来采用比较特殊的方法;一般缩略视频是相对于专题缩略视频而言的,它是一些视频片段的序列,用户可以通过播放这些相对短小的视频片段来了解整个视频的内容。与视频概要相比,缩略视频有其自身的优势,即缩略视频可能比视频概要中单纯的静止图像更加有意义,对用户而言,理解起来更加自然有趣,例如在纪录片中,视频的伴音就包含有重要的信息,因此,在很多情况下,以缩略视频作为摘要更加合适。

2. 视频概要的实现方法

视频概要是最能代表视频内容的静止图像集合,因此,关键帧的提取是视频概要实现的主要技术。目前概要生成的方法按帧、镜头、场景的视频层次结构划分,主要有基于镜头的概要生成方法和基于场景的概要生成方法两类。

(1) 基于镜头的概要生成方法。既然镜头被定义为一个连续的视频帧序列,那么在这个序列中就不存在场景或者摄像机运动的突变,因此一个很简单自然的方法就是把每个镜头的第一帧作为关键帧。如果镜头内的内容变化不大,则一帧关键帧就足够了;否则就应该提取多帧关键帧。但是,提取镜头中的哪些帧作为关键帧呢? 在目前计算机语义理解还很困难的情况下,大多以低层视觉特性(例如颜色、运动等)为衡量标准来抽取多帧关键帧。

(2) 基于场景的关键帧提取方法。对于基于镜头的关键帧提取方法,如果是长视频,那么将提取数以百计的关键帧,这样浏览起来不仅费时,而且低效。基于此原因,人们开始考虑基于更高一层的视频单元的关键帧提取法,称为基于场景的关键帧提取法。这里的场景比视频层次结构中的场景更广泛、更丰富,它可以是一幕情景、一个事件,甚至是整个视频序列。

除了以上谈到的用关键帧来构造视频概要的方法外,还有很多结合其他技术的视频摘要生成法,如马里兰大学把视频序列表示成高维特征空间的曲面来生成视频摘要。雅

典大学把模糊算法和遗传算法(genetic algorithxn,GA)运用到视频摘要中。此外还有结合小波变换、人脸探测等技术来提取关键帧的方法。

3. 缩略视频的实现方法

缩略视频有以下三种实现方法。

1）视频剪辑的实现方法

视频剪辑是一类比较特殊的视频摘要,它是原始视频中精彩场景的集合,但是并不包含故事的结局,通俗的称呼是片花。德国的曼海姆大学对剪辑视频曾做过研究,其研究焦点就是精彩场景的探测和选取。研究人员首先认为包含有强烈对比的前后帧可能包含有重要对象的重要事件;然后他们把表示整个视频段的基本颜色基调的场景也包括在视频摘要中;最后,把所有选取的场景按照时序组织起来,但是,在他们的研究项目中,由于研究人员对问题的复杂性尚考虑不够,所采用的算法还比较简单,因此效果有时候不是很好,还有待进一步提高。

2）专题缩略视频的实现方法

专题缩略视频是一种针对某一特定领域视频数据的缩略视频。对于专题缩略视频,一般可结合该领域的专题知识,采用特殊的方法来生成视频摘要。设计了一种专门针对该研究机构每周例会的视频摘要系统,即利用例会比较统一的履行程序,把低层的信号事件和高层的语义事件关联起来生成缩略视频。可见,专题缩略视频是从专题知识出发,更多的是采用基于模型而不是基于内容的方法来生成摘要。

3）一般缩略视频的实现方法

事实上,选取整个视频中最精彩的图像帧往往是由人主观确定的,而且如何把人的认识与计算机匹配起来是一件非常困难的事情。基于以上原因,目前缩略视频的重点集中在一般缩略视频的研究上。一般缩略视频实现的一个最直观的方法就是通过压缩原始视频来加速视频回放的速度。这种方法虽然有一定的效果,但是它存在压缩比的限制,因为这些压缩算法是依赖于语音速度的,如果压缩比过高,那么语音将无法理解。从目前视频摘要技术的发展来看,一般缩略视频的实现主要采用多特征融合的方法,也就是结合文本、音频和视频等媒体的特征来生成视频摘要。

综上所述,目前的视频摘要技术的研究重点主要集中在低层特征上,从而所形成的视频摘要不太符合人类的理解。在如何建立低层特征与高层语义概念的关联方面的研究目前还很少。在基于内容的视频检索中,视频摘要生成结果的好坏具有决定性的作用。因此,如何集成现有成熟技术到视频摘要系统中,使得视频分析与检索系统能够真正商业化应用,也是研究的重点问题之一。

9.10 视频语义检索模型

视频信息检索是多媒体领域的重要研究课题，是跨越图像处理、计算机视觉、模式识别、人工智能以及数据库技术等方面的交叉领域，是对文本、图像、声音等多种媒体形式的综合分析和查询。当前视频信息检索的研究主要集中在两大类：一类是基于视频低层特征的样例或样图查询（query by examples），另一类是基于视频描述信息的语义查询（query by keywords）。

第一类属于基于样本视频或图片的查询，是利用用户给出的查询样例，提取样例视频和数据库视频的低层物理特征，并根据一定的相似度度量，通过计算二者之间的相似度得到用户所需的查询结果。

第二类属于基于关键词的查询，是通过对视频库中的视频数据进行高层语义分析，通过用户提供的查询关键词对视频内容进行检索。

这两类视频检索方法分别从低层物理特征和高层语义特征两个方面，对视频内容进行分析和检索，是视频检索领域两个重要的研究方向。从 2001 年至今，诸如 CMU、IBM 等研究机构已相继提出了一些优秀的高层语义提取算法，并且取得了较好的研究成果。

视频语义检索模型主要组成模块包括底层特征提取模块、底层特征向高层语义映射模块、视频语义查询模块。其模型图如图 9-6 所示。

9.10.1 底层特征提取模块

该模块主要包括视频镜头检测、关键帧提取、特征提取三种关键技术，这三种关键技术在本章的前半部分进行了叙述，此处不再进行赘述。

9.10.2 底层特征向高层语义映射模块

底层特征空间包括视觉特征和非视觉特征，这些特征一般可以从视频数据中直接提取。语义概念空间对应于人们通常思维中的高级语义概念。从认知层次角度进行视频语义划分的语义概念，主要包括事件、场景地点和对象三类。但底层特征对用户不可见，只有将其映射到高层语义概念空间，才能使用户识别，它们之间无法直接用数学模型完成映射转换，这两个空间之间存在着难以直接跨越的语义鸿沟，如何解决语义鸿沟是视频语义检索研究的一个重点问题。

底层特征向高层语义映射模块主要是映射变换模型的构建，即语义概念分类模型的

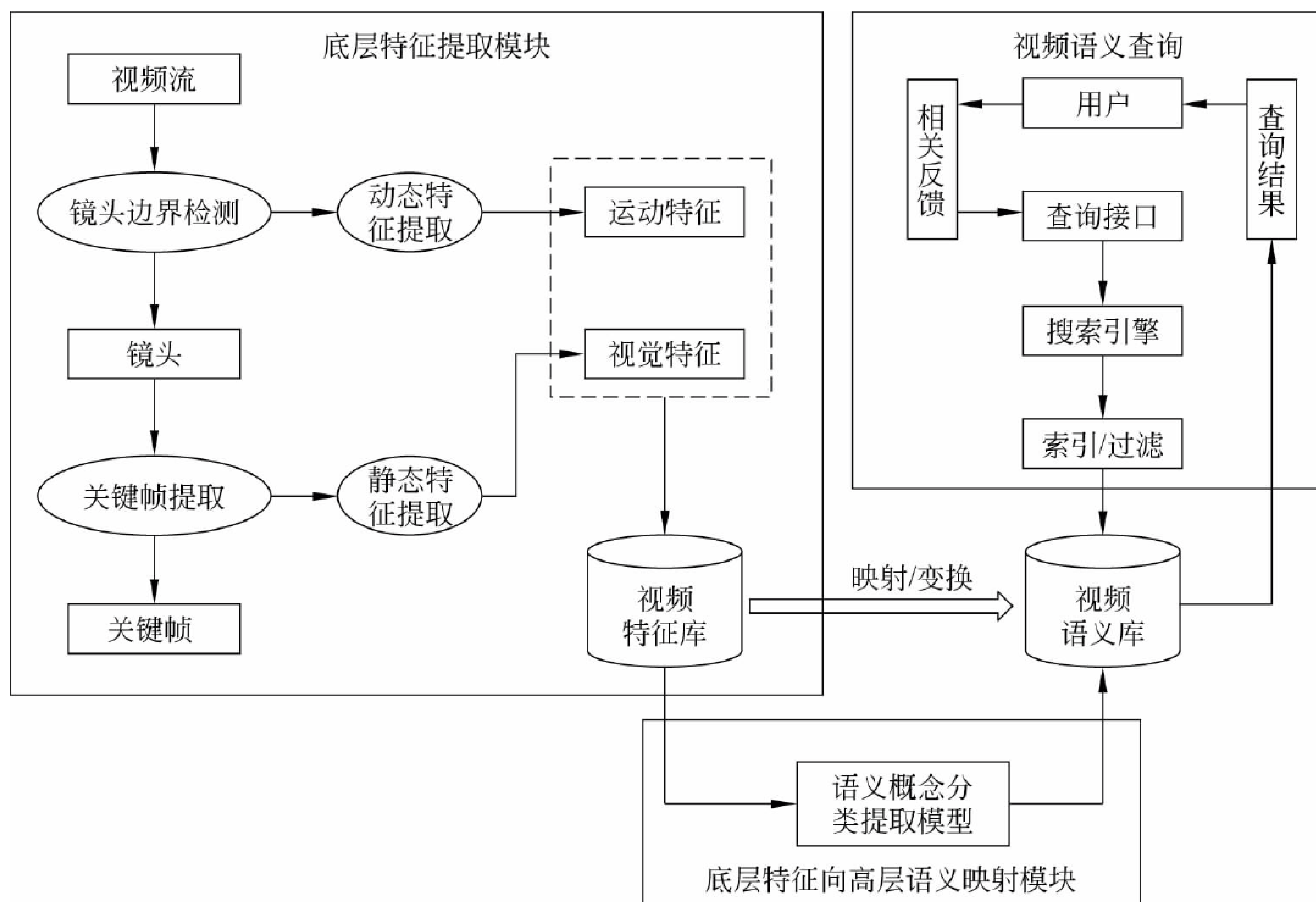


图 9-6　视频语义检索模型

构建。提取视频语义的主要方法包括概率统计方法、统计学习方法、基于规则推理的方法、结合特定领域的方法等。

(1) 概率统计方法。将视频语义对象提取看做是待提取视频语义对象的分类问题，利用模式分类方法来尝试跨越语义鸿沟。语义检索的随机方法关注的是模型概率特性，其核心思想是用随机数学方法来描述对象的不同特征并在此基础上建立多媒体概念模式分类器。随机模型中加入学习/识别模块，主要是为了能反映媒体内容本质的非确定性。

(2) 统计学习方法。基于支持向量机(support vector machine，SVM)的统计学习理论，建立在计算学习理论的结构风险最小化原则之上。其目的是在高维空间寻找一个超平面作为两类的分割，以保证最小的分类错误率。此类模型在只有小训练样例集的情况下，分类效果较好。先提取训练图像库的底层特征信息，然后利用 SVM 对所提取的特征进行训练，构造多分类器。在此基础上，利用分类器对测试图像自动分类，得到图像属于各个类别的概率，从而建立这些底层特征与视频类型之间的联系。

(3) 基于规则推理的方法。基于规则推理的方法考虑直接从系统外给定分类标准，

因此语义概念的种类固定，难以满意地描述视频内容中大量随机出现的语义概念。例如，通过分析足球视频的语义结构，按照足球比赛转播、视频编辑的一般规律，结合视频特征的时空关系，定义足球视频主要的语义规则，从而提出了足球视频语义事件的分析框架并结合基于专业知识的规则推理，达到有效分析足球视频语义的目的。

(4) 结合特定视频域。限定、缩小视频域 (narrowing the donmin)是目前跨越语义鸿沟的有效方法之一。限定特定的领域后，语义概念和事件的随机性就被缩小了，简化了底层和高层之间的语义映射关系。例如在影片语义分析领域，结合影片的特点只用四个视觉特征将电影分为悲剧、动作、戏剧和恐怖片几种类型，达到影片语义分类的目的。

上述这些方法在视频语义概念分类中虽有一定的应用但效果还不理想，有待于进一步完善与发展。而目前基于支持向量机(SVM)的方法在语义概念分类中显示出一定的优越性。视频语义查询模块使用户通过查询接口输入相应的查询语义，系统应能在视频语义库中进行信息匹配。并将查询结果返回用户。用户根据本次查询结果与自己期望结果间的相关性，向系统提交相关反馈信息。系统则根据用户的反馈来自动调整查询的内容继续检索，使查询结果向用户期望最佳“逼近”。

(5) 基于支持向量机(SVM)方法的语义概念分类模型。支持向量机是一种非常流行的学习机器，从模式识别领域的角度看，它是一个有监督学习的分类器。使用它分类需要先训练，再预测测试数据，向量是它的操作对象。根据向量在空间的分布，可以分为“可分数据-线形机器”、“不可分数据-线形机器”和“非线形机器”这三种情况。首先从最简单的线形可分的情况开始，再逐渐讨论 SVM 在其他两种情形下的发展和变化。

SVM 的思想总结为：一方面，有意使特征(核)空间的维数足够大，使得可以在这个空间建立超平面形式的决策面。为了得到好的泛化性能，通过对所建立的超平面添加一些特定的约束条件来控制 VC 维数，降低模型复杂性，这导致训练数据的一小部分被抽出来作为支持向量。另一方面，在高维空间的数值最优化受到维数灾难的影响，通过使用一个内积核的概念和求解在输入(数据)空间用形式约束最优化问题的对偶形式，避免了计算上的维数灾难问题。

9.10.3 视频语义查询模块

视频语义查询模块使用户通过查询接口输入相应的查询语义，系统能在视频语义库中进行信息匹配，并将查询结果返回用户。用户根据本次查询结果与自己期望结果间的相关性，向系统提交相关反馈信息。相关反馈在信息检索中是一种指导性学习方法，用以提高系统的检索能力。近几年，人们对相关反馈有了很深的研究，许多新颖的算法被提

出，主要有三类：第一类是权重调整算法；第二类是基于支持向量机的反馈方法，是在每次反馈过程中对用户标记的正例和反例样本进行学习，建立 SVM 分类器作为模型，并根据该模型进行检索；第三类是基于贝叶斯准则的相关反馈方法，其基本思想是根据用户反馈的信息进行统计判断。

9.10.4　语义词典的应用

在视频检索系统中，利用文本标注对图像进行检索是比较常用的方法，但一般的系统都是先对标注做简单的文字匹配，然后提交相应的结果。文本标注和用户输入二者文字不同，而语义一致，这种方法就无法检索到相应的内容，虽然有些系统能对这类同义词做例外处理，但却无法穷举所有的情况，更无法对更高层次的语义做检索。许多研究把语义词典引入到基于语义的视频检索中来，实现图像语义关键词的扩充，提高了检索的全面性。WordNet 是一个英文词汇的语义本体，它以认知同义词集合为单位来组织词语的关系。其中词语的关系包括上下位关系、整体部分关系、同义反义关系等。正是由于 WordNet 的这种构建方式，越来越多的研究者将其引入到了信息检索领域。描述了一个基于本体词汇的三维模型语义检索的方法，该方法首先对一个三维模型库的词汇进行语义上的扩充，然后基于关键词进行检索，而不是简单的文字匹配。

9.11　典型的视频检索系统

关于基于内容的视频分析与检索，已经取得了很多研究成果。目前国内外已研发了多个基于内容的视频检索系统，典型的视频检索系统主要有以下几种。

(1) Visual Seek 和 Web Seek 系统。Visual Seek 是一个通用的搜索引擎，是一个基于 Web 的图像/视频搜索工具。它主要是根据所检索图像中不同色块的空间关系进行相似匹配，另外也用到颜色、纹理等特征提取技术。Visual Seek 提供了多种查询方法：根据视觉特征、图像注释、草图和 Web 上搜索所有特有的图像 URL。

Web Seek 是一个专用的面向网络的搜索引擎。它的目的是在互联网上建立一个可视化对象的自动词典供用户查询。与 Visual Seek 一样，它也是采用多特征提取技术进行匹配，并提供基于注释和基于图像视觉信息的用户查询接口。

(2) VideoQ 系统。该系统允许用户通过大量的视觉特征和空间关系进行检索，其目的在于研究基于视频对象的视频内容进行所有视觉特征的检索。该系统的研究成果主要包括视频内物体的自动分割、自动追踪多检索对象、视频镜头自动分割等。扩充了传统的

关键字和主题导航的查询方法，允许用户使用视觉特征和时空关系来检索视频。由于视频经过分类，所以用户浏览镜头十分方便。

（3）Marvel 系统。该系统是一个多媒体分析和检索系统，由 IBM 研发中心开发。Marvel 的目的在于帮助广播公司、图书馆等媒体行业管理庞大且增长迅速的多媒体数据，使之更有效、更智能。Marvel 系统包括两部分：多媒体分析引擎和多媒体查询引擎。Marvel 技术使用了独特的方法对音频、视频、文本信息进行分析和理解，并对多媒体的内容自动地进行注释。

（4）MediaMill 搜索引擎。一个语义搜索引擎，包含了阿姆斯特丹大学在图像视频检索方面的最新成果。如颜色描述算子设计、压缩码本设计、社会标记（social tag）相关性分析等。

（5）Informedia 系统。卡内基·梅隆大学的 Informedia 数字视频图书馆系统，结合语音识别、视频分析和文本检索技术，支持 2000 小时的视频广播的检索；实现全内容的、基于知识的查询和检索。

（6）国内典型的视频检索系统主要有：Ifind 信息检索系统、NewVideoCAR 新闻节目浏览检索系统、MIRC 多媒体信息检索系统、清华大学开发的 TV-FI 视频节目管理系统、汉图智能分析与视频检索系统、千视通海量视频处理与检索系统、九凌视频分类检索系统等。

由此可以看出，目前国际上已经对视频分析技术进行了比较深入的研究，并已经取得了许多研究成果，但这些成果大多集中在对于一些底层结构和底层语义特征的分析方面，而对于高层结构和高层语义特征的研究还不成熟。

本 章 小 结

视频数据库既包含了视频数据本身的内容，也包含了不同视频数据间的关联数据。视频数据库系统的基础是视频数据模型，数据模型包括数据结构和操作。其中数据结构既要研究与数据本身内容相关的对象，也要研究描述不同视频数据间关系的对，而数据操作则只是数据的各种加工利用方法。

视频数据不仅数量大，结构复杂，数据冗余性突出，而且视频信息的丰富内容带来人们解释的多样性和模糊性。视频图像除了图像本身特有的冗余信息以外，还包括图像间的冗余信息，即相邻的视频图像往往具有相同或相似的空间和视觉特征分布。视频数据压缩较成熟的标准是 MPEG 系列标准。

基于内容的视频检索(content based video retrieval,CBVR)指根据视频的内容及上下文关系,对大规模视频数据库中的视频数据进行检索。基于内容的视频检索系统,先将视频流通过镜头边界检测分割为镜头,并在镜头内选择关键帧,再提取镜头的运动特征和关键帧的视觉特征,作为一种检索机制存入视频数据库,最后根据用户提交的查询,按一定特征进行视频检索,将检索结果按相似度呈现给用户,用户可以优化查询结果,系统会依用户意见灵活优化检索结果。

镜头是视频数据的基本单元,所以基于内容检索的视频处理,首先要把视频自动地分割为镜头,以作为基本的索引单元,这个过程就称为镜头边界的检测,也叫场景转换检测(scene change detection,SCD),它是实现基于内容的视频检索的第一步。通常的边缘检测方法是先通过边缘检测算子找到图像中可能的边缘点,再把这些点连接起来形成封闭的边界。基本的镜头边界检测算法有两类,一类是基于图像特征的非压缩域边界检测,另一类为基于编码信息的压缩域边界检测。

非压缩域的镜头分割常用方法有基于像素的方法、基于直方图的方法、基于块的方法、基于边缘改变比例的方法等。压缩域中镜头分割常用方法有基于 DCT 系数的方法、基于 DC 系数的方法、基于运动矢量和宏块预测信息的方法等。

一个镜头包含大量信息,在视频结构化的基础上,依据镜头内容的复杂程度选择一个或多个关键帧代表镜头的主要内容,因此关键帧(或关键帧序列)便成为对镜头内容进行表示的手段。关键帧提取方法主要分为两类:基于全图像序列的方法和基于压缩视频的方法,具体有基于镜头边界的方法、基于内容分析的方法、基于光流的运动分析法、基于聚类的方法、基于压缩视频的方法等。

视频数据的特征又分为静态特征和动态特征。静态特征的提取主要针对关键帧,可以采用图像特征提取方法,如提取颜色特征、纹理特征、形状和边缘特征等。获取视频运动特征的方法是运动估计(motion estimation),运动估计是指从当前帧图像中获取运动趋势和走向的过程,是数字视频稳像技术、视频压缩编码技术的核心步骤。对于不同的视频图像序列帧间的运动采用不同的变换模型,常用的三种模型有平移模型、相似模型和仿射模型。

运动估计的基本思想是将图像序列的每一帧分成许多互不重叠的宏块,并认为宏块内所有像素的位移量都相同,然后对每个宏块到参考帧某一给定特定搜索范围,根据一定的匹配准则找出与当前块最相似的块,即匹配块,匹配块与当前块的相对位移即为运动矢量。视频压缩的时候,只需保存运动矢量和残差数据就可以完全恢复出当前块。

视频图像序列的实质是灰度发生连续变化的一组图像,灰度投影法就是利用图像的

灰度分布变化特性获得图像的全局运动位移矢量，这与块匹配法利用单像素信息先获得小块的局部运动矢量后获得全局运动矢量不同。灰度投影法是利用图像序列的行列各自投影曲线做互相关处理，进而获得图像序列的全局运动矢量。

特征匹配的基本原理是：通过在参考帧中选取典型特征作为标识，并在当前帧中以一定的匹配准则进行搜索，以寻找对应的特征结构，从而获得图像序列的全局运动矢量。

光流是空间运动物体在观测成像面上的像素的运动速率分布，反映了在一定时间间隔内由运动所造成的图像变化。光流中既包括了被观察物体的动态行为信息，也包括了有关的结构信息。它利用图像序列的像素强度数据的时域变化和相关性来确定各自像素的位置的“运动”。

视频聚类是研究视频流中镜头之间的关系，也就是把内容相近的镜头重新组合在一起，用以描述视频中有意义的事件，或是为了缩小检索的范围，提高检索的效率。视频结构索引所要达到的目标就是能按需求随机定位到视频的某一帧。因此，只要能为视频数据里每一帧图像建立好索引信息，就可以在任何时候从该帧访问视频数据。

视频摘要就是对一个较长的视频文件的内容所进行的一个简短的小结。视频摘要是静止图像或者是运动图像的序列（这些图像序列可以附带音频也可以不带），这个序列比原始视频要短很多，但是这个序列应保留原始视频的基本内容，以便能够实现对原始视频进行快速浏览和检索。视频摘要就是通过对视频进行分析处理来自动生成紧凑的能够充分表现视频语义内容的静止或者运动的图像序列。

视频语义检索模型主要构成模块包括底层特征提取模块、底层特征向高层语义映射模块、视频语义查询模块。

本章思考与练习题

1. 视频数据至少有哪两个基本的层次结构？
2. 简述帧、镜头、关键帧与场景的各自含义。
3. 视频数据有哪些显著特点？
4. MPEG 的含义是什么？其主要作用是什么？MPEG 有何优点？
5. MPEG 的数据流包含哪三种成分？
6. 如何理解基于内容的视频检索的概念含义？
7. 简述基于内容的视频检索系统结构。
8. 简述视频镜头分割的含义。镜头边界检测算法有哪两类？

9．非压缩域的镜头分割方法突出有哪几种？

10．压缩域的镜头分割常用方法有哪些？

11．镜头切变和镜头渐变的概念含义？

12．检测镜头切变方法的基本思想是什么？

13．关键帧提取的基本原理是什么？

14．视频关键帧提取方法主要分为哪两类？请举例说明。

15．目前关键帧提取的主流技术是什么？如何理解其基本思想？

16．视频语义提取的含义有哪些？

17．视频数据的特征分为哪两类？

18．说明全局运动矢量的均值计算方法与权重值计算方法的原理。

19．常用的视频运动估计数学模型有哪些？

20．视频运动估计的基本思想是什么？

21．如何理解运动估计块匹配法的原理？

22．块匹配运动估计有哪些技术指标？

23．如何理解灰度投影法的含义？请举例说明。

24．特征匹配的基本原理是什么？有哪些基本逻辑步骤？

25．光流法的含义？有哪些光流的计算方法？

26．说明视频聚类的含义与基本思想。

27．视频摘要的概念含义是什么？

28．简述视频摘要和缩略视频的实现技术。

29．视频语义检索模型主要构成模块有哪些？简要说明各个模块的含义。

30．有哪些典型的视频检索系统？对其中一个视频检索系统的应用方法进行详细说明。

第 10 章　Web 信息搜索

Web 是 WWW(World Wide Web,万维网)的简称,它是 Internet 最基本、最广泛的应用服务,也是最主要的信息资源类型。在当今信息化社会,无论政府、企业还是个人对信息查询与获取都有强烈的需求,谁能更快更有效地获取最新、最准确和最全面的信息,谁就能在学习、生活或工作中取得优势。但是“信息越多等于没有信息”已成为人们的普遍共识,在海量信息中对于特定的信息需求而言,大量的垃圾信息会淹没所需信息。因此,Web 信息采集与搜索技术也就应运而生。对于信息用户而言,直接面对的 Web 信息获取工具就是网络搜索引擎,Google、Baidu 等搜索引擎是 Web 信息采集与搜索的典型代表。

10.1　搜索引擎概述

搜索引擎(search engine)是指根据一定的策略、运用特定的计算机程序搜集互联网上的信息,在对信息进行组织和处理后,为用户提供检索服务的网络系统。据统计,搜索引擎应用是位于电子邮件和社交网络工具之后的第三大互联网应用,成为人们获取 Internet 信息资源的重要工具和手段。

搜索引擎源于 1990 年由蒙特利尔大学 Alan Emtage 等三名学生发明的 Archie,它依靠脚本程序自动搜索并分析 FTP 服务器上的文件名信息,然后对其进行索引构建,用户必须输入精确的文件名进行搜索。Archie 是第一个自动索引互联网上匿名 FTP 网站文件的程序,但它还不是真正的搜索引擎。1994 年 4 月,斯坦福大学的两名博士生,美籍华人杨致远和 David Filo 共同创办了超级目录索引 Yahoo,由于它所收录的网站都附有简介信息,所以搜索效率明显提高。后来随着访问量和收录连接数的增长,Yahoo 开始支持简单的数据库搜索,但因为 Yahoo 的数据是手工输入的,所以也不能真正被归为现代搜索引擎范畴,事实上只是一个可搜索的目录。

现代意义上最早的搜索引擎出现于 1994 年 7 月,卡内基 · 梅隆大学的 Michael Mauldin 将 John Leavitt 的 Spider 程序接入到其索引程序中,创建了著名的 Lycos,它除了相关性排序外,还提供前缀匹配和字符相近限制,并第一个在搜索结果中使用网页自动

摘要,而它最大的优势是远胜过其他搜索引擎的数据量。此后,搜索引擎进入了高速发展时期,目前互联网的搜索引擎已达数百家,其检索的信息量也十分庞大。

10.1.1　搜索引擎基本结构

一般情况下,将搜索引擎分为采集器、索引器、检索器和用户接口四个部分,见图 10-1。

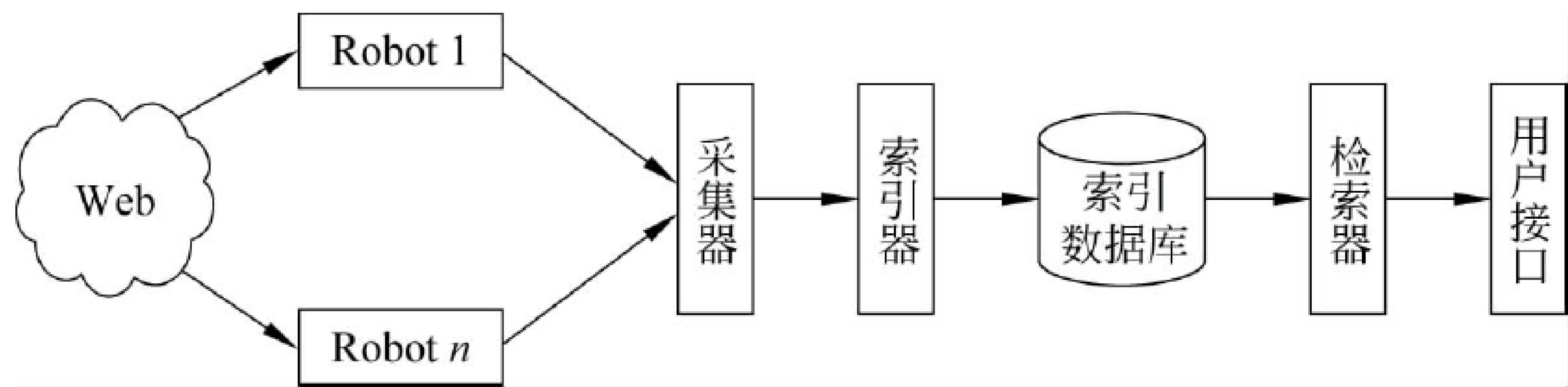


图 10-1　搜索引擎基本结构

(1) 采集器。采集器的核心就是网络蜘蛛(常常称为网络爬虫),它的主要作用就是按照预先设定好的算法,从网络上尽可能多地搜集相关的信息。不同的搜索引擎根据其搜索的主题的不同可以将网络蜘蛛的算法进行调整,以符合相关度采集的需要。同时,采集器还要定期进行更新,其目的就是对网络上已经消失或者过时的信息链接进行删除和更新。

(2) 索引器。索引器就是将采集器搜集来的链接或者信息进行分类,并按照一定的规则进行排列生成索引。索引器首先要抽取链接中的主题部分,将其作为索引项,并按照字顺或者数字顺序进行排列,生成索引表。由于大部分搜索引擎都是按累计词频来排列相关度的,加上中文词频的高频词往往又没有实际意义,所以索引表往往是按倒序排列的,又称为倒排索引。

(3) 检索器。检索器是承载用户接口与索引器的桥梁,用户将检索词提交给搜索引擎,用户接口将其传递给检索器,检索器根据用户的检索提问,将其规范化成主题词,并从索引数据库中查找相关的主题词,最后将查找好的链接或者信息提交给用户。检索器一般还要按相关性对结果进行排序,以利于将最相关的结果呈现给用户。

(4) 用户接口。用户接口是用户可以直接看得见的。无论传统搜索引擎,还是智能搜索引擎,它们的原理和结构基本相似,采集器、检索器和索引器都是在后台工作的,用户根本看不到它们是如何工作的,只有用户接口是用户交互的模块。用户接口设计得好与坏直接关系到搜索引擎的受欢迎程度,大部分搜索引擎的用户接口都是简约的、直观的和

方便应用的。用户接口要使用人机交互的理论和方法来实现,尽量符合人们的信息使用习惯。

10.1.2 传统搜索引擎基本类型

搜索引擎经历了近30年的快速发展,其形式和所专注的内容不尽相同。按照不同的分类标准,传统搜索引擎可以划分为不同的类别。

(1) 目录搜索引擎。目录搜索引擎也被我们称为网络目录,它是按照信息的主题进行分类存储和链接的一种简单直观的搜索引擎。它一般按照主题领域划分,每个主题又包括3～4层分支目录,我们通过这些目录来进行信息检索,对用户来说目录既方便又直观,有利于用户快速找到相关信息。网络目录通常采用网络信息分类法,依据网站性质或重点可以着重突出某些内容和信息,也可以根据实时情况增加某些目录。目录确定之后,然后将搜索来的信息分门别类地存储在目录之下,以供用户检索浏览,用户可沿着分类目录链接逐级浏览查找所需信息而不用关键词法进行查询。目录搜索引擎的缺点是:由于人们对主题的认识不同,造成对信息的分类也不同,这样就造成了大量不相关的信息存储在同一目录下,目录层次太少则造成信息检索的精度降低,目录层次太多会使信息检索检全率较低。

(2) 全文搜索引擎。全文搜索引擎就是利用“蜘蛛”(spider)或“机器人”(robot)搜集网络上的网页,然后将网页分类,组织到搜索引擎数据库中,并将每个网页进行全文标引。全文标引完成后,搜索引擎将标引过后的词句建立索引,形成索引数据库。当用户通过检索接口进行检索时,检索接口就将用户的关键词与索引数据库进行匹配,将匹配较高的网页和信息反馈给用户。

全文搜索引擎由于将全文进行标引,这样大大提高了检全率,只要是搜索引擎搜集到的网页,通过全文检索都能检索得到。因此对于谷歌和百度为代表的搜索引擎,用户就可以用同一检索词所搜集的网页数量来评价搜索引擎之间的检索优劣。但是全文搜索引擎致命的缺陷就是检索的相关度不高,许多与主题无关的词被当做关键词检索到。另外随着网页数量的增多,越来越多的不相关信息被检索到,真正与主题相关的信息却被深埋在信息海洋之中。

(3) 元搜索引擎。元搜索引擎不是一种有自己独立的结构或者特殊技术的搜索引擎,它是在检索时通过对其他独立的搜索引擎进行调用,并对搜索结果进行整合和优化的搜索引擎。元搜索引擎避免了用户在检索时频繁更换搜索引擎以期达到最相关的搜索结果的需求,用户不需要来回用相同的检索词在不同的搜索引擎之间查找比较,元搜索引擎

就可以对各个搜索引擎进行检索,并将结果提供给用户。元搜索引擎可以根据用户的检索提问,可以指定检索的顺序,控制检索时间,合理规范和整合检索结果,同时,也会自动处理检索过程中的重复、相同与雷同结果,以统一界面人性化显示检索结果。元搜索引擎没有建立独立的索引数据库,它只是一个对多个搜索引擎进行综合提问的检索接口。

(4) 集合式搜索引擎。集合式搜索引擎即将许多搜索引擎整合在一个单独的页面上,用户可以选择一个或者多个搜索引擎进行检索。当用户选择完搜索引擎之后,多个搜索引擎就同时开始检索,并将结果呈现给用户。集合式搜索引擎不能算做是真正的搜索引擎,它只提供一个有多个搜索引擎检索的界面,方便了用户选择搜索引擎。

(5) 垂直搜索引擎。垂直搜索引擎也称做主题搜索引擎或者专题搜索引擎。它是对网页库中的某类专门的信息进行一次整合,只关注某一领域或者某地域的信息,对这些信息存储和索引之后,用户就可以检索只涉及这一领域的信息。垂直搜索引擎与通用搜索引擎的最大区别就是:通用搜索引擎是面向所有用户的,而垂直搜索引擎是面向某一领域的用户的。生活休闲类搜索引擎是在 2006 年之后逐渐兴起的一类垂直搜索引擎,它主要搜集某个地域内生活休闲类信息,例如,酒店、道路、公交、商店、景点、娱乐、餐饮等信息,并按照用户所需地域自动将当地的生活信息提供给用户,极大地方便了用户的出行和旅游。

10.1.3　智能搜索引擎基本类型

智能搜索引擎与传统搜索引擎的结构原理有一定的区别,大部分智能搜索引擎是在传统搜索引擎基本结构的基础上,增加了相关技术或者相关系统优化原理而形成的综合检索系统。从信息的搜集到信息的组织与索引以及信息的检索与用户接口,智能搜索引擎在不断优化传统搜索引擎的各个方面。按照不同的分类,智能搜索引擎的结构也不尽相同,原理也有所差异。根据智能搜索引擎的分类和采用的相关技术,智能搜索引擎所呈现的特征也不尽相同。

1. 基于本体的智能搜索引擎

它设计的根本目的就是为了提高搜索引擎的准确性、语义性、个性化,同时利用智能化技术对搜索引擎的处理过程进行优化。基于本体的智能搜索引擎一般的信息处理过程主要如下:用户首先通过用户界面提出检索请求,检索器接受检索请求;然后本体编辑器对检索请求的格式进行规范,以符合本体的要求;接着将规范的请求格式提交给推理机,推理机依靠本体库的知识进行推理、判断和语义分析并最终得到准确的语义概念;最后利用准确的语义概念与用户偏好库同本体库中的概念与知识进行匹配,输出检索结果。基

于本体的智能搜索引擎主要结构模块有用户接口模块、搜索模块、检索模块、本体编辑器模块、推理机、本体库和用户偏好库，如图 10-2 所示。

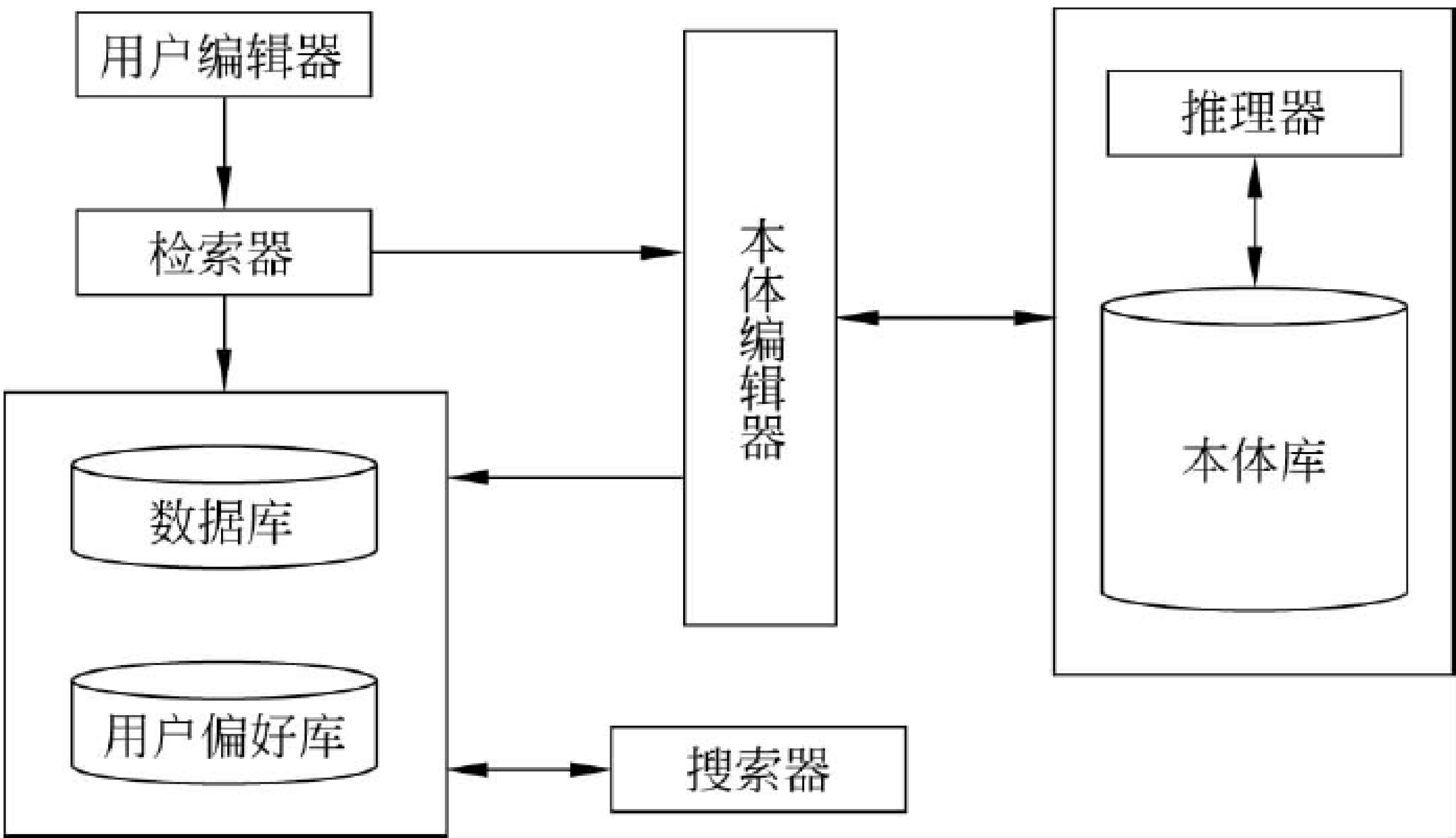


图 10-2　智能搜索引擎模块结构

（1）用户接口模块是搜索引擎与用户直接接触的模块，用户通过输入相关关键词或者相关概念来进行检索。用户不仅可以使用主题和关键词，如“×××大学”、“×××市政府”，也可以输入概念，如“番茄”、“计算机”等进行检索。当然，用户也可直接输入自然语言或者直接输入问题进行检索。

（2）搜索模块是传统搜索引擎的重要组成部分，同时也是智能搜索引擎的重要组成部分。但基于本体的智能搜索引擎的搜索模块是经过本体编辑器编辑过的检索请求，有针对性地对相关网站进行信息搜集，由于搜索器只对符合领域本体的文档进行搜集，这样很大程度上提高了信息相关度。同时，搜索器还参照用户偏好库的记录对相关信息进行筛选，能给用户提供更加准确的信息。

（3）检索模块首先要接受用户的检索请求，检索模块按照本体要求，将用户请求交给本体编辑器进行编辑，并将其转换成规定的格式提交给推理机。推理机经过相关推理，得到用户请求的相关本体，检索模块按照本体对数据库中的信息进行检索，查找匹配概念，最后将匹配结果再提交给用户界面。检索模块还有一个重要的作用就是将用户请求提交给用户偏好库，让用户偏好库将信息记录下来，这样可以及时更新用户偏好库，以供下次方便查找。

（4）本体编辑器是基于本体的智能搜索引擎特色模块之一，由于本体编辑器能将自然语言和概念进行编辑，有利于对用户请求的规范化处理，使检索请求更加准确，同时，本体编辑器要结合本体库中的本体，对用户偏好库中的信息进行规范，转换成本体所需要的

表达形式,使用户偏好库中的信息更加准确。

(5) 推理机是对经过本体编辑器编辑后的请求信息进行推理和判断,推理和判断的过程主要借鉴专家学者在思考问题时的推理和判断过程。经过推理机的推理和判断,再根据用户偏好库的用户个性特征,最终得到用户真正所需的信息,使信息检索的结果更加准确。

(6) 本体库是对领域内的相关知识按照本体内部的相关概念和规则进行规范化处理,使其在语义上准确地表达信息的概念及概念间的属性关系的数据库。本体库不仅表达领域信息的本身,同时也对领域信息资源的关系进行描述,进而形成知识网络。本体库的建设是由领域专家和本体专家共同设计和完成的,通过专家对领域知识进行总结和归纳,并构建本体之间的关系,最终形成本体库。

(7) 用户偏好库是用来存储用户经常查看的或者用户感兴趣的信息数据库。用户每次利用搜索引擎搜集信息,用户偏好库都对其进行记录和存储,并经过本体编辑器进行规范,当用户下次使用搜索引擎进行信息搜索的时候,搜索引擎自动访问用户偏好库,将用户感兴趣的信息主动提供给用户。

2. 基于知识库系统的智能搜索引擎

基于知识库的智能搜索引擎是智能搜索引擎中的一种,它利用知识库系统强大的理解能力和推导能力并运用人工智能技术,提高搜索引擎的智能性。它对知识有一定的理解与处理能力,可以实现同义词聚类、概念搜索、机器翻译等。主要模块结构有知识库系统、智能搜索器、索引器、检索器、结果反馈模块,如图10-3所示。

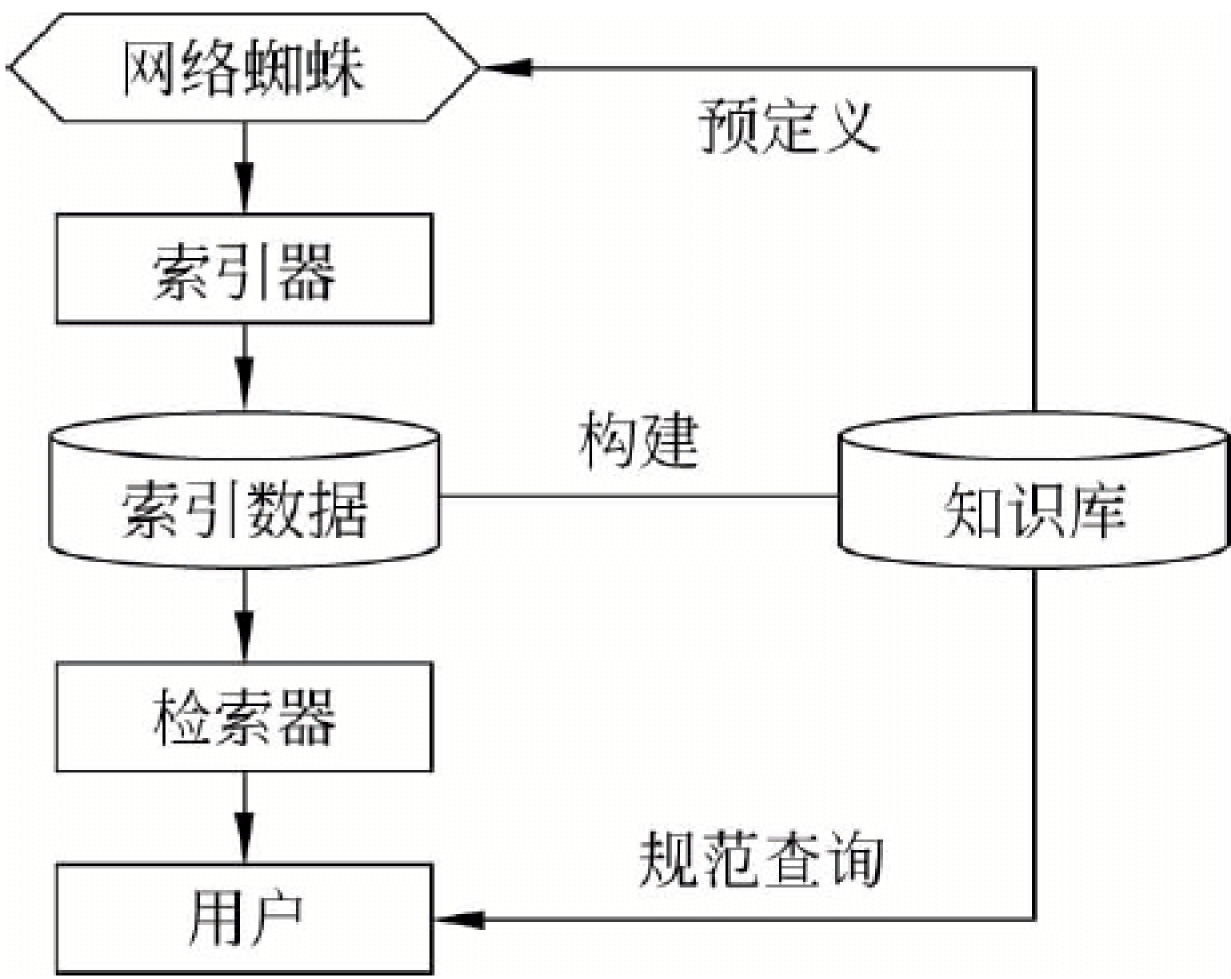


图10-3 基于知识库系统的智能搜索引擎模块结构

(1) 所谓知识,一般是经过人类利用归纳或者总结等方式加工整理而成的,是人们对现实世界客观的、正确的认识,这些认识对人类的发展具有重要的指导和引导作用。知识库系统是利用数据库存储知识,并设计相关算法按照一定的规则对知识进行推理,以便让机器更好地理解词语的意思,提高计算机的理解能力。知识库中存储的知识是程序在推理和解释的过程中所需要的知识,而不是向搜索引擎使用者提供的知识。知识库系统通过对用户输入的检索请求进行分析和推理,有助于检索系统理解用户的真正用途,使用户能获得高相关度的信息。知识库系统一般由知识库、推理机组成。推理机

负责对搜索请求进行分析和逻辑推理，它是知识库的核心组成部分，它利用推理规则、知识、专家词典等为基础，模仿人类推理问题的相关方法，最后得出相关结论。知识库中存储的知识大部分都是分层存储的，并以结构化的方式分布存储，这样有利于知识的发现和利用。

（2）信息检索主要的功能是进行检索的预处理，按照用户请求进行检索并对结果返回。为了使搜索引擎能理解用户的检索请求，必须对用户的查询词进行预处理，搜索引擎利用知识库系统的知识，对用户的请求规范化和精确化，进而得到用户请求的相关概念和语义概念。同时，搜索引擎还利用知识库系统，对查询的概念进行扩充和联想，使得检索结果能更加全面。查询预处理系统还将概念词返回给用户，以供用户选择更准确的检索词。然后，检索系统根据检索词从数据库中搜索相关信息，呈现给用户。

（3）索引器主要是用于自动标引和自动信息分类。知识库的重要组成部分就是概念，索引器依据知识库系统中的概念及概念间的关系进行标引和分类，索引器主要通过与知识库中相关概念进行比较，判断信息中的相关词汇是否与知识库中的概念相一致，通过匹配判断来对文档进行标引。

（4）智能搜索引擎的网络蜘蛛是智能化的网络蜘蛛，它主要依据知识库中的概念以及语义对相关的网络进行遍历，并将遍历结果提交给数据库。

（5）结果反馈主要是检索器将搜索结果提供给用户，同时需要用户对相关结果进行评价，以供搜索引擎进行统计和分析。当搜索引擎对大量的反馈结果进行统计和分析之后，就能自动识别用户的检索请求，或者主动提供相关词汇供用户选择。

3. 基于语义关联的智能搜索引擎

基于语义关联的智能搜索引擎是研究比较多的智能搜索引擎之一，它利用语义关联技术对搜索引擎进行设计和构造，以提高搜索引擎的智能化。领域本体是对某一学科内的概念明确的规范化说明。本体不仅反映领域内的概念，同时能表达概念之间的语义关系。所以领域内的语义关系一般通过本体来表示，通过本体并根据不同的搜索算法和匹配规则在文本中查找到相似概念，确定相关词汇，以此来提高搜索引擎对文本概念层次上的理解能力。基于语义关联的智能搜索引擎是通过预先定义学科本体即确定学科领域概念来实现对整个搜索过程进行改造，以达到真正理解用户的请求。其工作步骤如下：①用户使用语义关联编辑器对学科的相关概念进行创建和编辑，用户还可以增加和删除预定义的关联词族。②通过对学科概念的编辑确定本体的主题。③利用预定义的本体词族抽取 Web 文档的相关主题。④将抽取到的主题词进行保存，一般将主题词保存为属性索引。⑤用户利用检索系统接口对所需信息进行检索，输入检索词，获取检索结果。

基于语义关联的智能搜索引擎不仅提高了搜索引擎的信息检索准确性，同时也提高了资源间的关联程度和资源的利用深度。基于语义关联的智能搜索引擎的主要模块为：概念词表定义模块、概念词表导出模块、概念索引模块、概念检索模块、概念导航模块、结果反馈模块等。

（1）概念词表定义模块。通过概念词表编辑器，用户可以自定义概念词汇、建立词汇间的关联。概念词表的定义首先要进行核心概念词的定义，核心概念词是相关词族的标识词汇，只有对核心概念词定义以后，才能定义其从属词汇和概念间的关联。概念词表模块输入的是用户希望定义的概念系统的相关词汇，输出的是经过定义后的核心概念词、从属概念词汇和概念关系。

（2）概念词表导出模块。它将用户定义的概念词表以结构化的格式导出，以供搜索引擎模块在检索时使用，其根本的作用是将用户自定义的概念词汇传递给搜索引擎。

（3）概念索引模块。它通过对 Web 文本进行概念提取，并将提取后的概念词建立成索引文件，然后建立概念索引数据库。概念索引模块主要有以下工作：首先将输入的文本拆分为单个词组或词汇，并根据概念词表中的概念进行分类，使拆分的词汇尽可能地表达文本的内容。接着将拆分好的词汇与概念词表进行匹配，并将匹配成功的概念词汇输出。最后，根据输出的概念词汇，生成概念索引文件，概念索引包括核心概念词串、文件位置链接和相关描述等内容。概念索引能最大限度地反映文本的内容，提高文本资源的信息利用程度。

（4）概念检索模块。它是对概念索引进行遍历和搜索的模块。用户通过概念导航模块、概念范围收缩及关键词检索来实现概念检索。它还可以对用户输入的词语进行规范化处理，并将处理结果返回给用户，以供用户修改和选择检索词。概念检索模块还要将用户规范化的检索词与概念索引进行匹配，并将匹配结果如：文件链接、文本概念相关词汇、文本概念词、原始词等传递给用户。

（5）概念导航模块。它将概念分层级展示给用户，用户可以通过预先查看概念目录的内容选择自己需要的节点直接查询。这种形式有利于用户方便快捷地检索，是重要的辅助子系统。

（6）结果反馈模块。它是将用户查询的结果以及用户选择的结果进行记录，通过记录来反映搜索结果与用户需求是否相关，以此来提高搜索引擎的检索结果相关性，同时，结果反馈也有利于搜索引擎掌握用户的检索兴趣与检索习惯，对搜索引擎的个性化检索有很大帮助。

10.2 搜索引擎主要支撑技术

搜索引擎技术原理的种类较多,主要因其应用的信息采集算法原理和索引技术的不同而不同。目前,搜索引擎的主要支撑技术有分词技术、网络蜘蛛、索引技术、词频相关指数、自动推理技术、本体知识系统、专家系统等类型。

10.2.1 分词技术

分词技术是中文搜索引擎特有的一种技术,评价中文搜索引擎的优劣的一个重要指标就是分词技术。在汉语的语法和句子中,词汇以字为单位,两个字或者多个字构成一个词,词与词之间不像英文由空格分开,各个词之间没有空格,几乎无法将词语分别开来。因此需要分词,就是将由多个连续的字组成的关键词或句子重新按指定的算法分割成若干个有独立含义的字或词。中文词汇的组合非常灵活多变,组合后的词语意思也不尽相同,很容易对文字的理解产生歧义。如,对关键词“北京的大学”,可以切分为“北京/的/大学”,由于“的”属于助词,往往又将其切分为“北京大学”。由于切分方法的不同,可能造成几种不同的切分结果,返回的查询结果也会迥然不同。因此,分词的准确性将直接决定搜索引擎的查询结果。目前中文分词的算法主要有三大类:基于字典的分词技术、基于统计的分词技术和基于规则的分词技术。其中基于字典的中文分词技术占主导地位。基于字典的算法主要有两种:正向最大匹配法和逆向最大匹配法。

(1)正向最大匹配法。正向最大匹配法就是将段落分成句子,将句子分成词语,即将大化小,将小短语进行分解。它的分词方法是:将分词词典中最长的词语取出来,我们假设其长度为 L,即该词语包含 L 个汉字,然后从文章中第一个字开始,取前 L 个汉字与词典相配,如果匹配成功,则这个词就被切分开来,作为一个词语。如果匹配不成功,则从下一个汉字开始,重新匹配全文。如果按此方法匹配完成以后,则将 L 个汉字去掉最后一个字,即现在要匹配词典中 $L-1$ 个词,按照前面的方法,以此类推,直到将所有的词语切分出来。最终,将整篇文章或者整个段落切分完成。

(2)逆向最大匹配法。逆向最大匹配法和正向最大匹配法类似,只是在匹配的时候是从信息最末端开始匹配,匹配结束后去掉的不是最后面的字,而是最前面的字,其使用的分词词典也与正向最大匹配法有所不同。逆向最大匹配法是从被匹配信息的最后面开始扫描匹配,即从末端最后一个词开始。取词典中 L 长度汉字的词语开始匹配,若匹配成功则作为切分词,若匹配不成功,则去掉 L 长度汉字的最前面的词即 $L-1$ 长度的词继续

匹配。从去掉词的前后可以看出,逆向最大匹配法与正向最大匹配法所使用的词典也是不相同的。它使用的分词词典是逆序词典,其中的每个词条都按逆序存放。在实际处理时,将文档按照一定的规则进行倒排处理,生成逆序文档,然后根据逆序词典,对逆序文档用正向最大匹配法处理。根据数据显示,逆向最大匹配法相比正向最大匹配法效果要好得多,其误差也比较小。由于最大匹配法是一种基于分词词典的机械分词法,使分词结果不能很好地体现文档的语义特征,另外,它必须依赖词典进行分词,所以在实际使用时,难免会造成一些分词错误。一般情况下,我们在分词的时候都采用正向匹配法和逆向匹配法相结合的方法,通过两者的结合,可以达到理想的结果。

10.2.2　网络蜘蛛

网络蜘蛛又称之为 Spider 或者 Robot,其具有独立的工作能力与决策能力,它是通过网页的链接地址来寻找网页的,它在网络上查找相关信息,并将搜集到的信息返回给服务器。网络蜘蛛的本质是人造的一段程序代码,由于网络蜘蛛的目的就是永不停歇地抓取网络资源,就像我们常常见到的蜘蛛一样,在自己编织的网上爬来爬去,因此我们形象地将之称为"蜘蛛"或"爬虫"。网络蜘蛛从网站某一个页面的首页开始,读取网页的内容,找到在网页中的其他链接地址,然后通过这些链接地址寻找下一个网页,这样一直循环下去,直到把这个网站所有的网页都抓取完为止。如果把整个互联网当成一个网站,那么网络蜘蛛就可以用这个原理把互联网上所有的网页都抓取下来。

网络蜘蛛有很多种,不同的搜索引擎一般都会有其专门的网络蜘蛛程序。它们一般由不同的脚本程序编制而成,可以利用不同的编程语言来设计网络蜘蛛。

10.2.3　索引技术

索引即我们通常所说的按照一定的顺序将索引项目进行排列的一种方法。搜索引擎的索引技术是搜索引擎的一项重要技术,它关系到搜索引擎结构的构造以及检索结果相关度的高低排序。搜索引擎一般按词频排列,特别是中文搜索引擎,由于常用的无实际意义的助词出现的次数比较多,但与主题又不相关,所以搜索引擎往往使用倒排索引。倒排索引常被称为反向索引,它是一种索引方法,一般用于全文检索时,指引词或字在数据库中存储的位置,通过这种一一对应的关系可以很快查找到相关主题词的位置。倒排索引通常有两种形式:一种是记录的水平反向索引,用于引用词语的列表;另一种是单词的水平反向索引,它包含所有单词在记录中的位置。单词的反向索引其兼容性比较好,可以提供短语搜索,可以更好地反映记录的主题,但单词记录数量较大,需要的存储空间也比较

大,所以需要的资源耗费也比较多。后者的形式提供了更多的兼容性,但是需要更多的时间和空间来创建。

10.2.4 词频相关指数

词频指的是在一篇文档或者记录中某个词语出现的总次数。某一词语出现的次数越多,代表该文档或者记录与该词语的主题越相关。通过对词频的统计可以确定文档或者记录的主题词语。但是只靠单一的词频累加方案往往又不能很好地反映主题,因此搜索引擎要利用一定方法来规范词频,以利于更加准确地找到主题词语。单文本词频指数和逆文本频率指数是文档资源的两个重要指数,它们是搜索引擎用来进行词语加权的两种重要方法,通过对单文本词频和逆文本词频的运算可以排除经常用到的无实际意义的词汇,能将高频词汇与主题词汇进行高相关度匹配。

10.2.5 自动推理技术

推理是指从已知的判断和条件下,推论出新的判断或者新结论的一种逻辑思维形式。推理是人们解决问题的一种常用方法,它是依靠人们对相关知识的掌握,并根据事物之间的联系来进行处理问题的一种方法。自动推理是人们利用计算机模仿人们推理问题的过程与步骤而自动得到解决问题的一种技术。自动推理主要由程序推导、程序结果证明、专家系统等相关部分组成。

程序推导主要涉及计算机算法,程序设计者根据人们推理问题的过程,设计出相关算法来模仿人类的推导过程,并用机器语言实现推理过程,这样计算机就能根据相关的前提条件自动进行推理。

程序结果证明就是人们证明定理的过程。通过一定的程序和算法加以形式化,使计算机能自动实现对推理结果的证明,这样有利于推理结果的精确性,防止错误的推理和不符合常识的推理的出现。

专家系统是对推理进行控制和判断的系统,通过专家系统存储的知识和逻辑判断能让推理过程更加智能化,推理结果更加合理和准确。

智能搜索引擎的自动推理主要依靠本体库或知识库中的本体或知识进行推理,通过知识的概念及概念间的关系理解,并采用相关算法自动推理出用户检索请求的相关概念和联想词,提高搜索的相关度。自动推理还结合用户相关反馈或者用户偏好库中的信息进行推理,以便准确地理解用户的检索请求。

10.2.6　本体知识系统

本体本身是一个哲学概念,后来被计算机科学引入,用来将领域内的各种概念及概念之间的关系准确地、形式化地表现出来。通过这种表示可以准确地获得概念之间的语义关系。关于本体的定义有许多种,但被人们广泛接受的是:本体是概念化的明确的规范说明。概念化主要是对客观事物的抽象说明,其表达的基本意义独立于具体的外部环境。明确化要求概念必须被准确地定义,并尽可能将概念规范地表示出来。本体所反映的知识是大家共同认可的,是领域内专家和普通用户广泛认可的概念集合。本体能表示领域内的概念及概念之间的关系,本体能对用户的搜索请求从语义方面去理解,消除歧义现象,所以通过本体能够准确地反映用户的真实信息需求。

知识系统是 20 世纪 70 年代被提出来的一种用来存储知识的系统。知识系统与数据库系统有很大的不同,数据库系统存储的主要是无序的数据或者是按照一定的规则排序的数据,这些数据的语义和它们之间的关系都不能被表示出来。知识系统存储的往往是领域的相关知识,并利用知识之间的关系建立一定的体系结构,形成知识系统。知识系统不仅仅存储知识,还可以根据知识进行相关的推理和演绎,具备一定的智能化性能。人们利用知识库系统可以来进行问题的求解,提高计算机的理解能力,使计算机能像人类一样思考和解决问题,而不是简单的机器翻译与理解。

由于本体可以有效地表达和查询知识,可以消除同义词之间的语义歧义,还可以支持语义发现,自动进行语义化的匹配和组合,因此可以利用本体来构建知识系统。本体知识系统提高了知识的利用深度,有利于对隐性知识的发现和获取,有利于知识的共享和创新。本体知识系统主要有以下功能:能识别多种表示语言形式和存储形式;能进行本体学习、本体映射、本体自动合并等相关操作;能支持本体的可扩展性和一致性,对本体的多个版本进行兼容化管理。

由于本体知识系统具有较强的语义理解能力和自动推理能力,对信息的处理和利用超出了数据库系统,能从知识的角度来管理和操作信息,使信息检索上升为知识的检索,因此本体知识系统是智能搜索引擎的一项重要技术。智能搜索引擎通过利用本体系统,有助于对关联词的理解,并结合自动推理技术,为用户提供联想词提示。通过本体知识系统可以实现对文本的智能分词,提高分词技术的水平,本体知识系统还可以对用户的检索请求与检索结果相关性进行总结,获得词汇的词频与用户请求之间的关联,这样就解决了单单依靠词频统计来确定词汇的相关度,提高了分词效果。

10.2.7 专家系统

专家系统事实上是一类智能的计算机程序系统，这类系统关注的重要内容就是专家的知识和经验。它通过利用计算机程序来模仿专家解决复杂问题的思维模式和过程，使计算机达到与专家同水平的解决问题能力，以提高计算机的智能化水平。因此要想构建一个专家系统必须要拥有相关领域的大量专家知识，同时要能模仿专家的思维模式，才能达到专家级别的解决问题的能力。

专家系统作为一个智能系统，构建起来就如同一项巨大的工程一样，因此专家系统也称做知识工程。专家系统通常由人机交互界面、知识获取、知识库、推理机、解释器、综合数据库组成。

人机交互界面是用户与专家系统进行交流的界面，用户通过人机交互界面输入相关信息和提问，专家系统接受用户的提问和信息，并经过专家系统的处理最终将结果提供给用户。

知识获取模块的功能主要是建立、修改和扩充知识库。它从专家的头脑中或者各种知识源那里获取知识，并将其转换成一定的格式存储到知识库中。知识的获取可以通过计算机自动获取，也可以通过人工的识别和分类进行获取，知识获取有利于专家系统知识库的更新，提高专家系统解决问题的能力。

专家系统是通过推理机来分析和推理的，依靠这种方法来解决用户提出的问题，知识库是专家系统的心脏，推理机是专家系统的大脑。它根据知识的语义，按照一定的逻辑算法，找到相关知识并提供给用户。推理机的算法和程序与知识库的内容是相互独立的，推理机的程序与知识库的具体内容无关，这样的好处是：如果知识库进行更新或改动就不会对推理机的推理算法和程序造成影响。

解释器是用来向用户解释说明专家系统求解问题的过程，让用户明白专家系统是如何工作的，并对用户的提问进行回答。解释器提高了专家系统的透明性，能让用户明白专家系统正在做什么和为什么要这样做，用户也是通过解释器来认识专家系统的工作原理的。

综合数据库是用来反映专家系统对用户请求的求解状态集合的数据库，也可以将综合数据库称为动态库。它存放的是系统运行过程中产生的各种信息，包括系统需要的数据源、用户请求、推理的中间结果、推理过程等。综合数据库中由各种事实、命题和关系组成的状态是推理机选用知识的依据。

将专家系统应用到搜索引擎，提高了搜索引擎人机交互功能，可以帮助智能搜索引擎

总结用户的兴趣,并主动将信息推送给用户。依靠专家系统,可以实现用户的个性化搜索,建立用户的个人门户。专家系统记录用户的个人喜好,并跟踪用户的搜索轨迹,建立符合用户个性化需求的信息服务平台。

10.3 Web采集

10.3.1 Web采集概述

随着互联网的迅速发展,人们接触最多的信息是以Web页面形式存在的。我们面临一个信息爆炸、信息困扰的时代。面对互联网上兼具多样性和复杂性的海量信息,仅仅依靠人工搜集与整理来有效跟踪最新信息动态显然是不科学的和低效的,也不能满足实际需要。于是人们开始探索新的信息获取方式,Web信息采集技术应运而生。

随着网络应用的深化和技术的发展,Web正由以搜索引擎为主的单纯检索服务向着信息转播、个人代理、个性化主动服务等领域全方位拓展。作为这些服务系统的重要基础和支撑,Web信息采集的任务也越来越艰巨,被广泛应用于搜索引擎检索、站点结构分析、页面有效性分析、Web图进化、内容安全检测、用户兴趣挖掘以及个性化信息获取等多种服务和研究当中。Web采集是从Web中收集网页的过程,这些网页用于索引从而为搜索引擎奠定基础。采集的目标是尽可能高效地采集更多数目的有用页面,并同时获得连接这些页面的链接结构。

10.3.2 采集器的功能与特点

Web采集器的功能特点可以分为两类:一类是采集器所必须提供的功能特点,另一类是采集器应该提供的功能特点。

采集器所必须提供的功能特点包括以下两点。

(1) 鲁棒性。Web中有些服务器会制造采集器陷阱(spider traps),这些陷阱服务器实际上是Web页面的生成器,它能在某个域下生成无数网页,从而使采集器陷入到一个无限的采集循环中去。采集器必须要能从这类陷阱中跳出来,尽管这些陷阱倒不一定都是恶意的。

(2) 完整性。Web服务器具有一些隐式或显式的策略来控制采集器访问它们的频率,设计采集器时必须符合完整性的访问采集策略。

采集器应该提供的功能特点包括以下六点。

(1) 分布式。采集器应该可以在多机上分布式运行。

(2) 规模可扩展性。在增加额外的机器和带宽的情况下，采集器的架构应允许实现采集率的提高。

(3) 性能和效率。采集器应能够充分利用不同的系统资源，包括处理器、存储器和网络带宽等。

(4) 质量。在应答用户查询需求时，大部分 Web 网页的质量都较差，因此采集器应优先考虑抓取有用的网页。

(5) 新鲜度。在很多应用中，采集器都处于连续工作状态，也就是说它应该要对原来抓取的网页进行更新。只有这样一个搜索引擎才能保证其索引包含索引网页的较新版本。对于这种连续式采集来说，采集器应能够以接近网页的频率来采集网页。

(6) 功能可扩展性。采集器的设计要能支持在很多方面方便地进行功能扩展，比如可以处理新的数据格式、新的抓取协议等，这就要求采集器的构架要高度模块化并具有充分的扩展接口。

10.3.3 Web 采集

目前，在 Internet 的各种应用中，以 Web 应用最为普及，发展速度尤为迅速，Web 上的信息资源也急剧增加。Web 资源的异构性、开放性和广泛分布性等特点，使用户在获取自己需要的信息资源时面临很大的困难。搜索引擎的出现为解决这一问题提供了重要途径，它也逐渐成为用户在 Web 上获取信息的主要工具。

信息采集指通过 Web 页面之间的链接关系从 Web 上自动地获取页面信息，并且随着链接不断向整个 Web 扩展的过程。任何超文本采集器(不论是面向 Web、内网还是其他的超文本文档集)的基本处理如下：首先，设定一个或者多个 URL 为采集的种子集合(seed set)；接着从种子集合中选择一个 URL 进行采集；然后对采集到的页面进行分析，并抽取出页面中的文本和链接(每个链接都链向其他的 URL)。抽取出的文本输入文本索引器，而抽取出的 URL 则加入到待采集 URL 池(URL frontier，以下简称 URL 池)中，任何时候 URL 池中放的都是所有待采集网页的 URL。实现这一过程主要是由 Web 信息采集器(Web Crawler)来完成的。Web Crawler 也常称做 Web Spider、Web Robot 或 Web Worm。简单地讲，它主要是指这样一个程序，从一个初始的 URL 集出发，将这些 URL 全部放入到一个有序的待采集队列里，而采集器从这个队列里按顺序取出 URL，通过 Web 上的协议获取 URL 所指向的页面，然后从这些已获取的页面中提取出新的 URL，并将它们继续放入到待采集队列里，然后重复上面的过程，直到采集器根据自己的策略停止采集。对于有些采集器，到此就算完结了，而对于另一些采集器，它还要将采集

到的页面数据和相关数据存储、索引并在此基础上对内容进行分析。

(1) 采集器架构。一个简单的采集器由多个模块构成，如图 10-4 所示，其中包括五种模块。

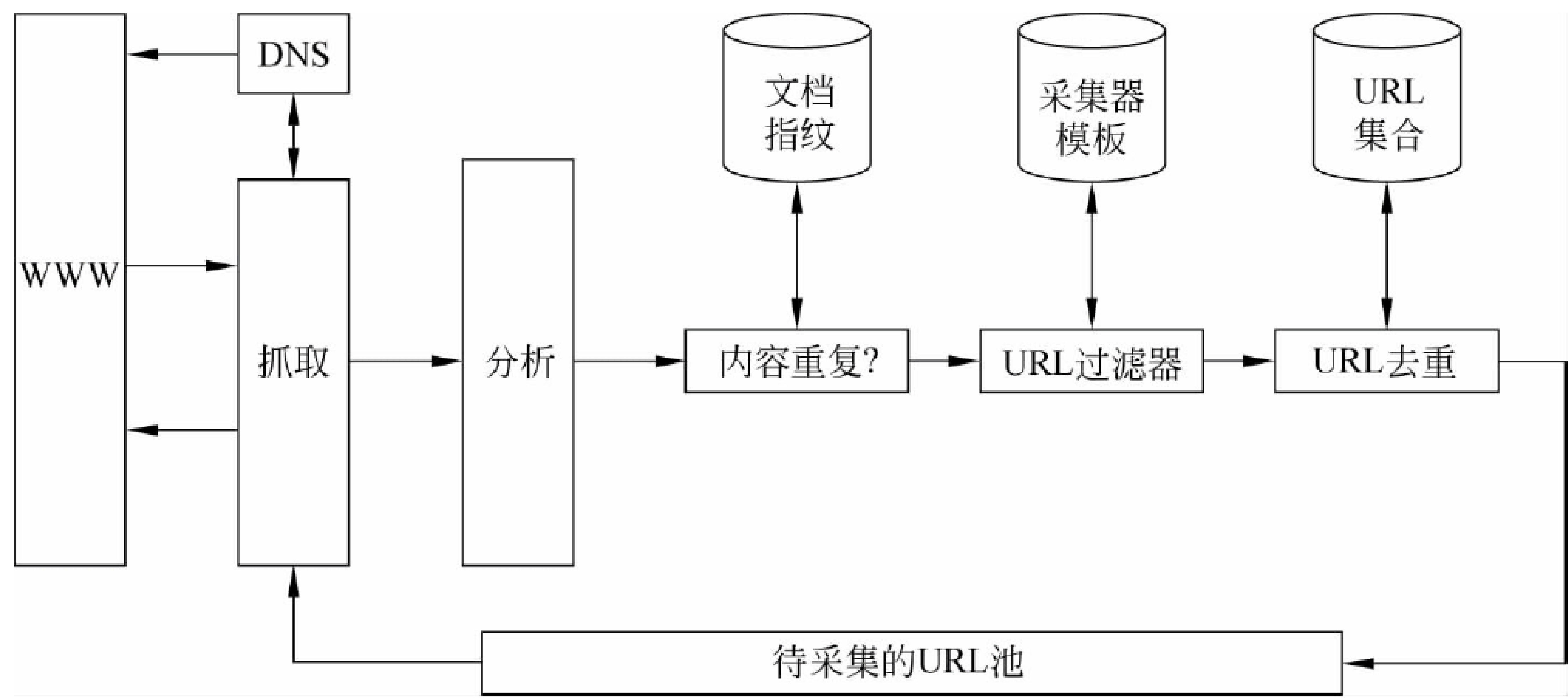


图 10-4　采集器的基本框架

① 待采集 URL 池。它包含了当前待采集的 URL(在连续采集中，某个已经采集过的 URL 可能还会放回到该采集池中以便进行重新采集)。

② 域名(DNS)解析模块。它在 URL 抓取网页时用于确定其对应的 Web 服务器的 IP 地址。

③ 抓取模块。利用 http 协议返回某个 URL 对应的网页。

④ 分析模块。从采集到的网页中抽取文本及链接。

⑤ URL 去重模块。确定某个抽取出的链接是否已在 URL 池中或者最近是否已抓取过。

(2) URL 的采集流程。主题信息采集模块负责从 URL 开始从 Internet 上获取信息，并对采集回来的页面进行处理。MRobot(制造网络机器人)负责从 Internet 上获取网页并进行处理，包括文档类型过滤、分析并提取链接、获取网页内容信息、对文档的文本内容进行关键词分析并形成网页数据库。

基于鱼群算法，同时结合首页关联技术、页面内容预测技术等各种主题采集策略的 MRobot 工作方式如下：

(1) 从初始 URL 队列(其初始值为预定的种子站点 URL)列表中获得一个 URL 请求页面。种子站点的建立采用了人工预选技术，具体来说就是运用 Google 和百度从互联

网上搜集一些国内外比较知名的制造资源网站以及一些制造行业的企业网站，再咨询业内的一些专家、教授来确定这些初始种子站点。以机械行业为例：机械行业主题的初始种子网站就是采用国内一些知名的机械类网站的网页作为初始种子（例如中国机械网）。它们在实际应用中得到了国内企业界的支持以及国内互联网行业的一致认可，具有广泛的知名度及很高的权威性。同时后续种子的添加过程中通过咨询业内的专家和学者不断地对种子网页进行更新和完善。

（2）分析获取的页面，提取超链接和页面内容信息。对种子站点内部链接直接插入待处理 URL 队列头部，将站外链接 URL 插入待处理 URL 队列最后端。

（3）对非种子站点的 URL，提取其首页分析其主题相关性。如果相关，则按照和种子站点的页面相同的处理方式，否则直接丢弃，对整个站点都不再采集。

（4）提取获取页面的内容信息，将结果添加到数据库。

（5）将 URL 加入已处理 URL 列表并获取下一个待处理的 URL，如此不断循环直到待处理 URL 列表为空。

具体来说，MRobot 采集流程如图 10-5 所示。为提高信息采集的效率，采用了多线程的技术同时对多个 URL 进行处理。网络机器能够自动地访问网络上数百上千的 Web 服务器站点。

10.3.4 域名解析

人们习惯记忆域名，例如桂林电子科技大学网站的域名是 www.guet.edu.cn，但计算机间互相只认 IP 地址，例如桂林电子科技大学网站的 IP 地址是 202.193.64.56，域名与 IP 地址之间是一一对应的，它们之间的转换工作称为域名解析，域名解析需要由专门的域名解析服务器来完成，整个过程是自动进行的。当网站设计完成后上传到虚拟主机时，可以直接在浏览器中输入 IP 地址浏览网站，也可以输入域名查询网站，虽然得出的内容是一样的，但是调用的过程不一样，输入 IP 地址是直接从主机上调用内容，输入域名是通过域名解析服务器指向对应主机的 IP 地址，再从主机调用网站的内容。

1. 树状结构的域名空间

为便于管理，Internet 中的域名采用层次结构，并用域名空间来描述，如图 10-6 所示。在域名空间中，把名字定义到一棵倒置的树形结构中（类似家谱树），树的每一级定义了域名层次的每一级。

树状层次结构上的每一个节点都有一个域名，每一个域名都由该节点向上读到根节点，通常根节点的标号为空。DNS 要求每个节点其下的子节点应具有不同的标号，因此

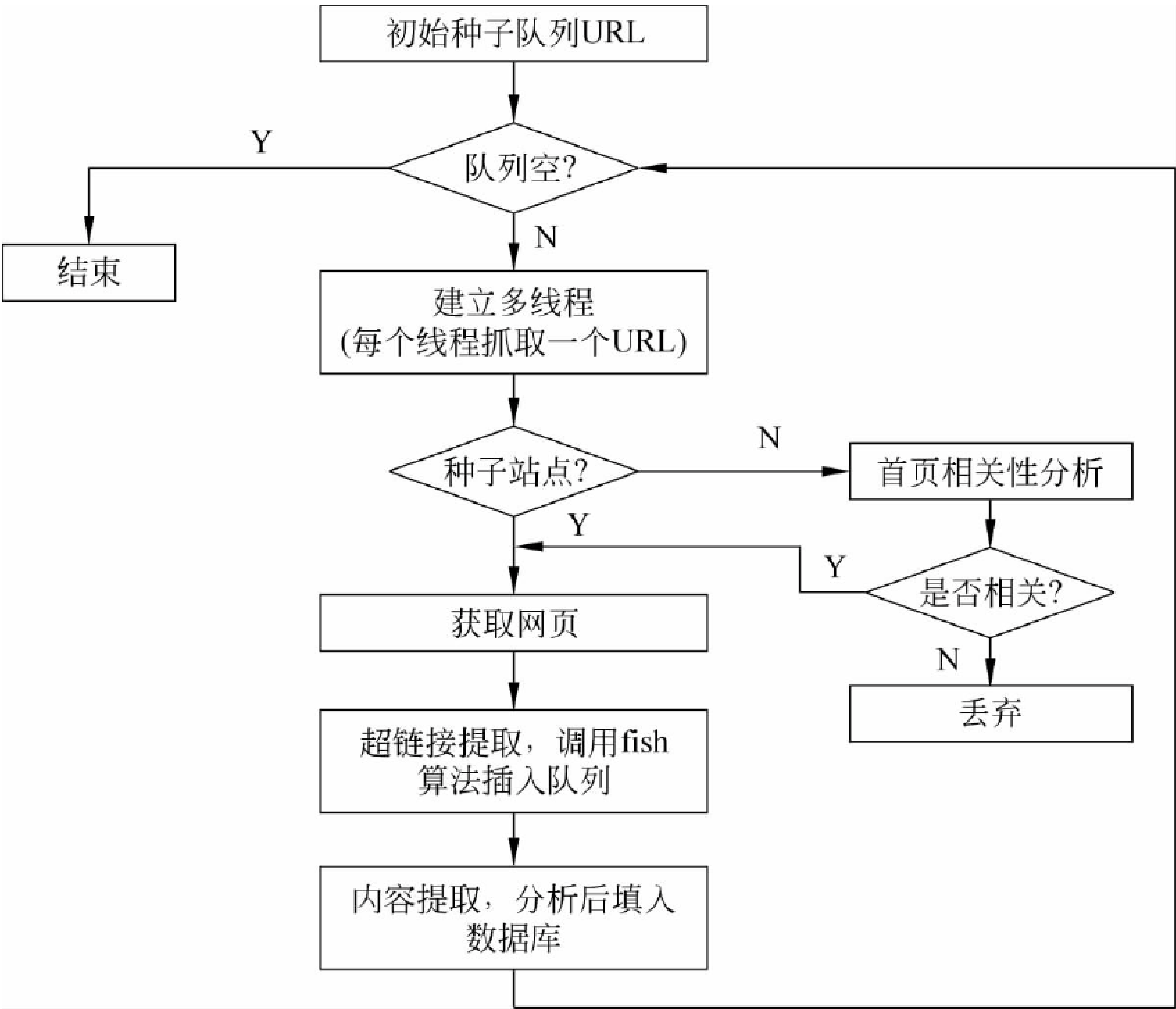


图 10-5　MRobot 主题信息采集流程

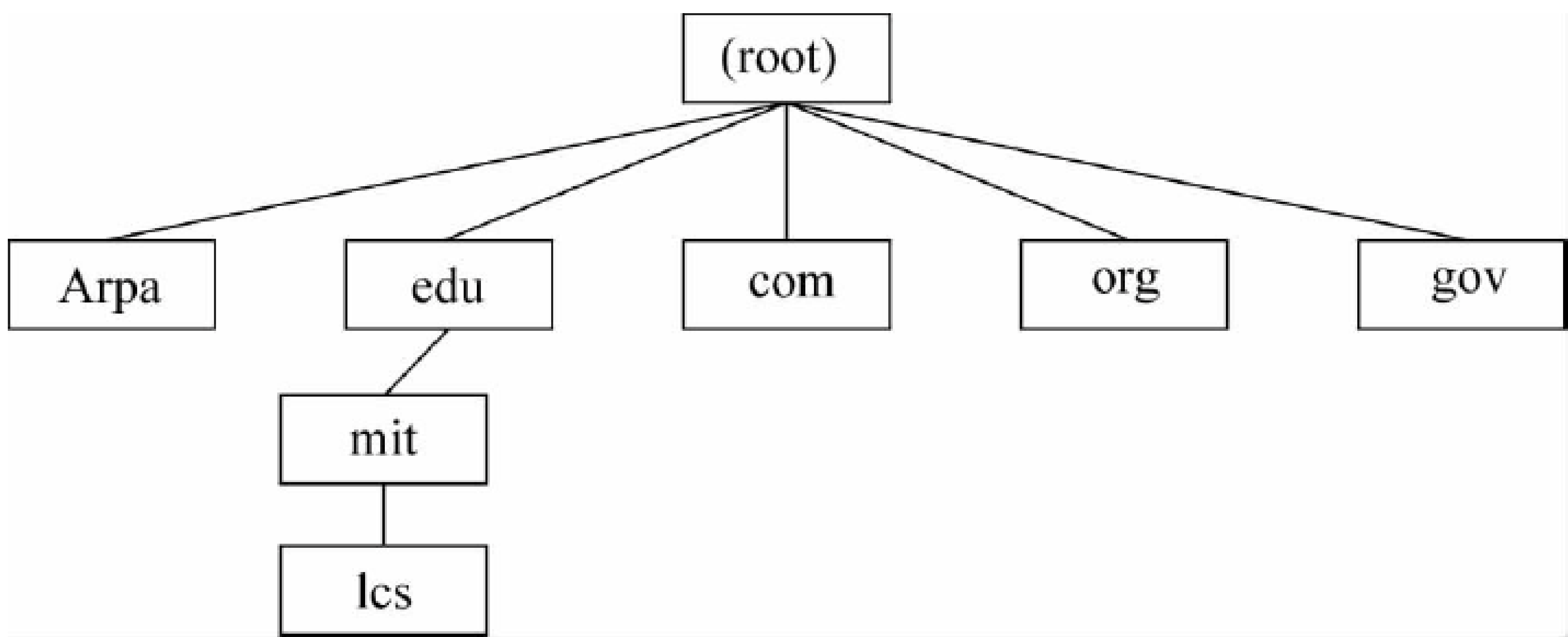


图 10-6　域名空间的层次结构

这种树状结构保证了域名的唯一性。

2. 地址解析

利用上面的层次结构，地址映射就可以分布到所有 DNS 服务器上了，这就为分布式数据库提供了依据。当主机发出它的 DNS 请求时，该请求首先被发往本地 DNS 服务器，

如有需要,本地 DNS 服务器将该请求转发到 DNS 层次结构中去,实现地址映射。在这种机制下,就涉及 DNS 客户机与 DNS 服务器以及各服务器之间的通信。这里我们通常采用的查询方式是递归查询和迭代查询。

递归查询。所谓递归查询,是指接受查询请求的第一个 DNS 服务器必须对请求进行处理并予以响应。假设主机 www. guet. edu. cn 要请求 www. baidu. com 的 IP 地址,本地服务器将该查询转发到上级服务器,上级服务器再将该查询转发到更高级的根服务器,直到查询被解析出来或出错为止。

迭代查询。如果客户机没有要求递归查询,则可以按迭代方式映射。迭代过程中的每一次查询请求都由客户机发出,对查询的每一次响应也直接返回给客户机,如果没有解析到 IP 地址,返回的内容将包括下一个逻辑上更近的 DNS 服务器的 IP 地址,客户机根据这个 IP 地址继续查询,直到返回的是最终结果或者出错。理论上,DNS 查询既可以是递归的也可以是迭代的。在实际应用中,通常把递归查询和迭代查询结合起来使用。

3. DNS 高速缓存

在信息量巨大的 Internet 中,网络带宽、服务器负载等是一直面临并努力优化的问题。设想一下,如果 DNS 服务器每次收到对本地之外的连接请求时都进行逐层查询,那么这些应用带来的网络通信量将相当巨大,并且带来额外的时延。为解决这一问题,DNS 设置了高速缓存,和很多方面一样,它采用冗余技术。DNS 缓存原理很简单:当 DNS 服务器收到一个 DNS 回答时,它就将映射信息缓存在本地存储器上,在下一次收到查询请求时,DNS 服务器就首先检查本地缓存,如果在本地缓存中存在所需信息,它就直接从缓存中取出信息回答客户机的请求,如果所需信息不在缓存中,则再发出进一步的查询。

每个 Web 服务器(实际上每个连入 Internet 的主机)都有一个唯一的 IP 地址。在 DNS 解析过程中,需要进行 IP 地址转换的程序(这里指搜索引擎采集器)会联系一个 DNS 服务器来返回 IP 地址。众所周知,DNS 解析在 Web 采集中是一个"瓶颈"。由于域名服务本身就是分布式的,所以 DNS 解析可能包括多个请求在 Internet 上的往返过程,这通常需要数秒甚至更多的时间。这样,就会给每秒获取数百网页的采集目标造成极大困难。一个常规的措施就是引入缓存机制。然而,遵循采集中的完整性要求往往又会限制缓存的命中率。DNS 解析还存在另外一个严重的困难,采集器的开发者往往使用标准库(这个库可能被开发采集器的任何一个人使用)来实现 DNS 解析功能,为了避免这个问题,大部分 Web 采集器都会采用自己的 DNS 解析器。

10.3.5 待采集 URL 池

在每个节点上,采集进程或其他采集进程的主机分割器会将 URL 放入本节点的 URL 池中。该采集池会维护一系列 URL,并在采集线程需要寻找 URL 时,以某种次序将 URL 输出。采集池中 URL 的输出次序必须要考虑到两个重要的方面。

第一,频繁改变的高质量网页应该优先考虑频繁采集。因此,网页的优先级应该是其变化率和质量的函数(可以采用某些合理的质量估计方法)。由于大量作弊网页在每次抓取时几乎完全改变,所以同时考虑变化率和质量这两者是十分必要的。

第二,要考虑完整性问题。我们必须避免在很短的时间间隔内反复访问同一主机。由于很多 URL 会链向同一主机的其他 URL,因此会产生互相引用的局部效应,因此,如果不进行控制,那么在很短时间内访问同一主机的可能性很大。所以,如果 URL 池的实现中只使用简单的优先级队列,就会造成对某个主机的突发性高频访问。甚至即使在我们限制在任何时刻最多只有一个采集线程访问某个主机的情况下,上述突发高频访问仍然有可能发生。一个普遍使用的启发式策略是,在对某个主机发送连续的两次抓取请求之间插入一个时间间隔,它要比最近一次从该主机抓取网页所需的时间高一个数量级。图 10-7 给出了 URL 池的一个实现示意图,它支持优先级处理并遵循完整性访问原则。其目标是为了保证:①在任一时刻只有一个连接对主机开放;②在连续两次主机请求之间,需要等待数秒;③高优先级网页优先采集。

URL 池通常采用基于层次语义的 URL 排序算法,逻辑如下。

(1) 输入目标主题,初始 URL 值,阈值 H。

(2) 根据目标主题,从领域概念树中获取知识路径"knowledge-path"。

(3) 按照"knowledge-path"构造主题层次,最小层号为语义不相关层,最大层号为目标主题层。

(4) 对各主题层(语义不相关层,层 0 除外)训练对应的一个分类器。

(5) 初始化一个 URL 等待队列(UrlQueue)。

(6) 初始 URL 进入 URL 等待队列(UrlQueue)。

(7) 提取 URL 等待队列的队首 URL 元素,爬取 URL 指向的 Web 文档 d。

(8) 基于层次语义的 Web 文档分类。将爬取过来的 Web 文档 d 分配至与它最相似的主题层,并赋予 Web 文档 d 的层次语义度量。

(9) 析取文档 d 中的 URL 链接,由链接信息库中链接状态过滤掉已被爬取和出错的 URL 链接。

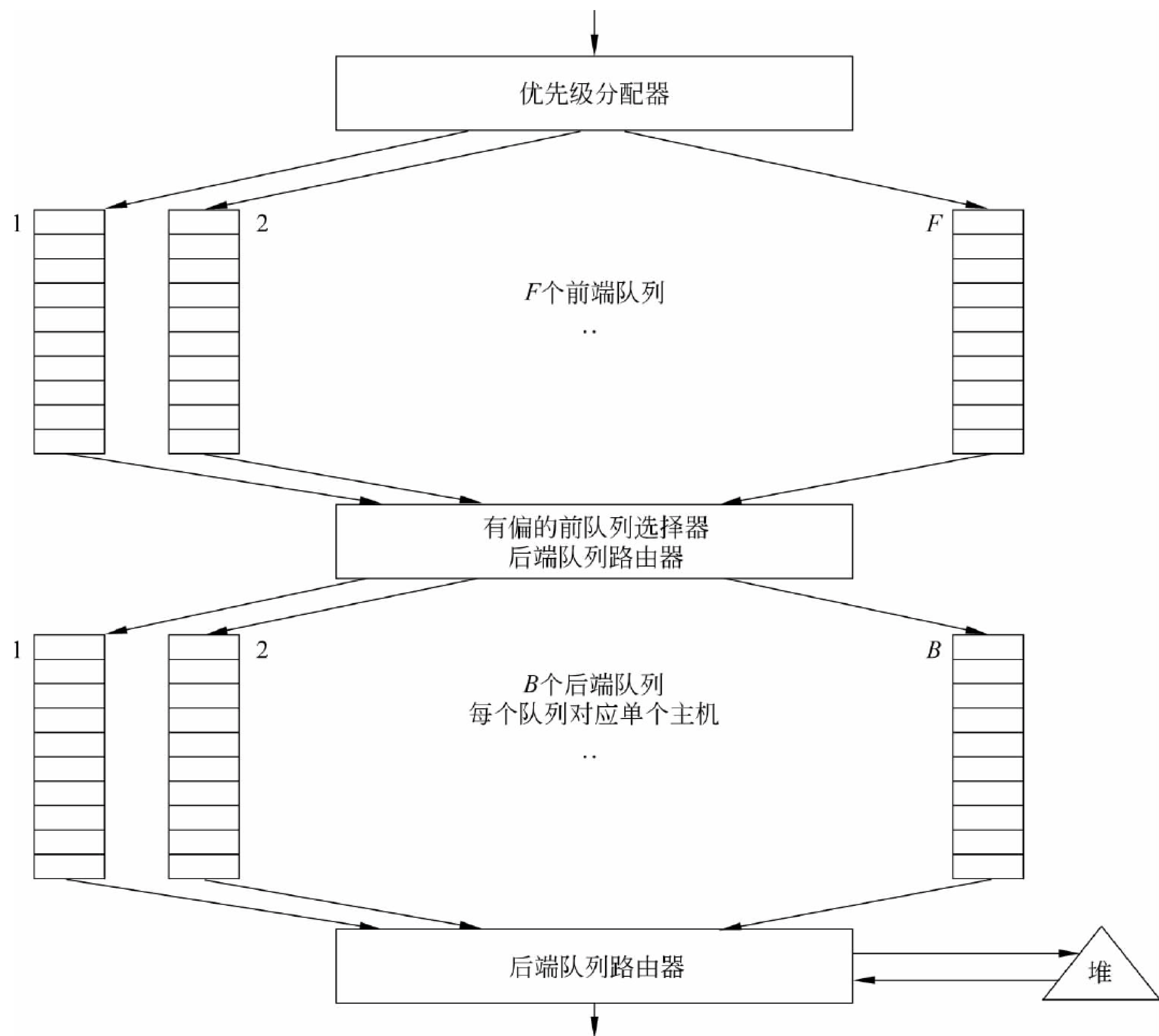


图 10-7　URL 池示意图

(10) 基于层次语义的 URL 链接排序。对 URL 等待队列(UrlQueue)中的候选 URL 链接按层次语义组合排序度量排序。

(11) 若 URL 等待队列不为空,则转(7),否则结束。

10.3.6　分布式索引

在信息爆炸式增长的今天,Google、百度等知名搜索引擎已经被广泛使用,它为信息收集带来了便利,然而在这背后支撑搜索引擎体系架构的一项重要工作基础是索引。当需要搜索大量的网页文档,并且想找出包含一个指定的词或短语的文档时,编写程序的一

个简单方法是针对给定的词或短语进行顺序扫描每个文档。这个方法有很多缺点，最明显的是不适合大量的文档或者文档非常巨大的情况。索引因此而产生，为了能够提高效率，先将文档转化为一个可以进行快速搜索的格式，避免传统缓慢的顺序扫描过程，这个转化称为建立索引。所以，可以把索引简单地理解为一个可以快速随机访问存在其内部词的数据结构。而如今数据的存储已经不只是一个文档、一台计算机，网络将存储的范围扩展到了全世界，要在如此海量的数据中快速检索更加需要新型索引的支持——分布式索引。

分布式信息检索是指由检索代理程序将检索任务同时提交给网络上的多个主机，由位于这些主机上的检索程序分别独立检索并将检索结果返回到检索代理程序，经过整理后显示给用户。

分布式信息检索由各种分布式 Web 服务器执行具体任务，虽然它们的工作原理不尽相同，但要解决的基本问题却大体一致：一是要有某种机制把客户的请求分派到各个成员服务器上，二是要有一个算法来指导请求分派以保证各个成员服务器的负载均衡，三是要在各个成员服务器恰当地复制和分布 Web 站点内容以维护其一致性并保证成员服务器的存取效率。

分布式索引构建方法是 Map Reduce 的一个应用。Map Reduce 是一个通用的分布式计算架构，它面向大规模计算机集群而设计。集群的关键是利用价格低廉的通用计算机（称为节点，node）来解决大型的计算问题，这些计算机都采用通用的标准部件（处理器、内存和磁盘），而不是像超级计算机那样采用专用硬件。尽管在这样的一个计算机集群当中包含成百上千台计算机，但每台计算机都有可能在任意时刻失效。因此，要保障分布式索引构建过程的鲁棒性，就必须把整个任务分成易分配的子任务块，并在节点失效时能够重新分配。集群中的主控节点（master node）负责处理在工作节点上的分配和重分配任务。Map Reduce 中的 Map 阶段和 Reduce 阶段将计算任务划分成子任务块，以便每个工作节点在短时间内快速处理。图 10-8 给出了 Map Reduce 的具体逻辑步骤。

一般来说，Map Reduce 会通过“键-值对”（key-value pair）的转换处理，将一个大型的计算问题转化成较小的子问题。在索引构建中，“键-值对”的形式就是词项 ID 与文档 ID 匹配。在分布式索引的构建过程中，从词项到其 ID 的映射同样要分布式进行，因此分布式的索引构建方法要比单机上的索引构建方法复杂得多。一种简单的解决方法就是维护一张高频词到其 ID 的映射表并将它复制到所有节点计算机上，对低频词则直接使用词项本身而不是其 ID，所有节点都共享一致的词项到其 ID 的映射表。

Map Reduce 的 Map 阶段将输入的数据片映射成“键-值对”，这个 Map 分别对应于相

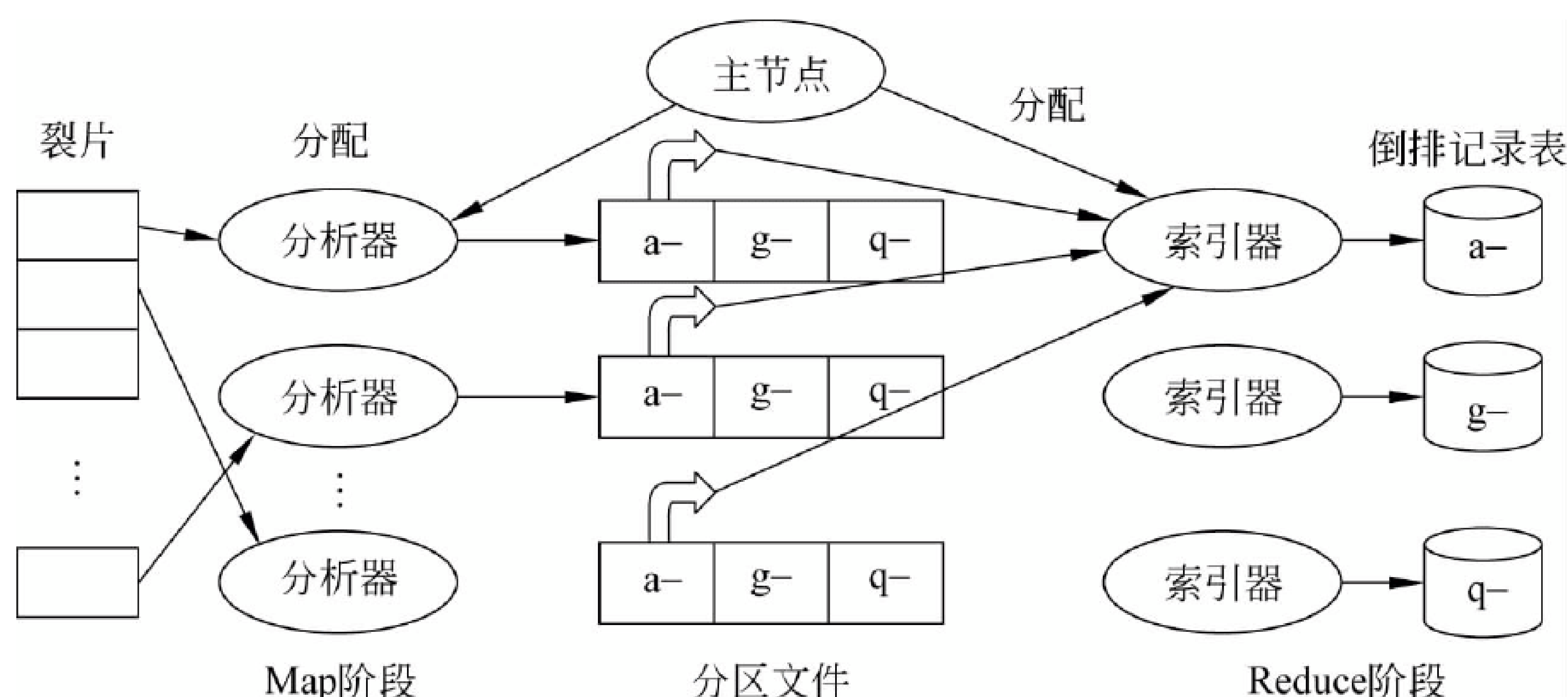


图 10-8　一个使用 Map Reduce 进行分布式索引构建的例子

应算法中的分析任务，因此也将执行 Map 过程的机器称为分析器(parser)。每个分析器将输出结果存在本地的中间文件(也称为分区文件，segment file)中。在 Reduce 阶段，我们想将同一键(词项 ID)的所有值(文档 ID)集中存储，以便快速读取和处理。

实现时，将所有的键按照词项区间划分成段，并将属于每个段的“键-值对”写入各自分区文档即可。图 10-8 中，所有的词项按照首字母来分成三段：a～f、g～p 及 q～z。词项的分割方法由运行索引系统的用户来定义。每个分析器各自写相应的分区文档，每个分区文档对应一个词项区间。因此，在整个系统中，每个词项区间会对应 r 个分区文档，其中，r 是分析器的个数。假设对于 a～f 分区，有三个 a～f 分区文件，它们分别对应三个分析器。

Reduce 阶段由倒排器(inverter)来完成，主控节点将每个词项分区分配给不同的倒排器，并在倒排器失效或者变慢的时候将在其上处理的词项分区进行重新分配。最后，每个键对应的所有值要进行排序并写到最终的排序倒排记录表(图中以“倒排记录表”来表示)中。需要指出的是，图 10-8 中每个倒排记录当中还包括词项频率，针对 a～f 分区处理的数据流如图 10-8 所示。到这里为止，整个倒排索引的构建才宣告完成。

分析器和倒排器并不一定是不同的机器，主控节点发现空闲的机器后会给它分配新的任务。同一台机器在 Map 阶段中可以作为分析器，而在 Reduce 阶段也可以作为倒排器。另外，索引构建的同时，机器上往往也在同时运行其他任务，所以在做分析器和倒排器之外，一台机器也可能运行采集程序或者其他不相关的任务。为了尽量减少在倒排器对数据进行 Reduce 之前的读写时间，每个分析器都将其分区文档写到本地磁盘。在 Reduce 阶段，主控节点会通知倒排器与之相关的分区文件的位置(例如，词项 a～f 分区对

应的六个分区文档)。

在每个分析器上,由于与某个特定倒排器相关的数据已经被分析器写入一个单独的分区文档中,所以每个分区文档仅需要一次顺序读取过程。这种设置方法可以使索引时所需的网络通信开销最小。图 10-9 给出了 Map Reduce 的通用函数构架。由于输入和输出通常都是“键-值对”列表本身,所以多个 Map Reduce 任务能够串行执行。实际上,这正是 Google 索引系统的设计方案。

Map和Reduce函数的构架
Map:输入　→list(k, v)
Reduce:(k, list(v))　→输出
索引构建中上述构架的实例化
Map: Web文档集　→list(词项ID,文档ID)
Reduce:(〈文档ID_1,list(doc ID)〉,〈文档ID_2,list(doc ID)〉, …)
→ (倒排记录表1,倒排记录表2,…)
索引构建的一个例子
Map:d_2: C died.d_1: C came, C c' e d.
→ (〈C,d_2〉 〈died,d_2〉,〈C,d_1〉 〈came,(d_1) 〈C,d_1〉,〈c' e d,(d_1)〉)
Reduce:(〈C.(d_1,d_2,d_3)〉, 〈died,(d_2)〉,〈came,(d_1) 〉,〈c' e d,(d_1)〉)
→ (〈C,(d_1:2,d_2:1)〉,〈died,(d_2:1〉,〈came,(d_1)〉,〈c' e d,(d_1:1)〉)

图 10-9　Map Reduce 中的 Map 和 Reduce 函数

10.3.7　连接服务器

自 20 世纪 90 年代以来,Internet 的应用在全球范围内得到了迅猛发展,一方面微处理器性能得到了很大的提高,Internet 基础设施也在不断提升,越来越多的计算设备接入到 Internet 中;另一方面,Web 应用正成为 Internet 上最重要的一种应用。据统计,Web 信息流量已经占到了整个 Internet 信息量的 80%以上,而且,越来越多的应用开始采用基于 Web 的 B/S 服务模式。

由于某些原因,Web 搜索引擎需要一个连接服务器(connectivity server)来支持 Web 图连接查询(connectivity query)的快速处理。典型的连接查询包括“给定的 ULR 被哪些 ULR 所指向”及“给定 URL 指向了哪些 URL”等。为此,我们在内存中存储了从 ULR 到链出及 URL 到链入的映射表。

假定整个 Web 包含 40 亿网页,每个网页有 10 个链接指向其他网页(这种情况称为链出)。在最简单的形式下,对每个链接的首尾两端(链接源和链接目标),分别采用 32 比特位或者说 4 个字节来描述,于是总共需要 $4\times10^9\times10\times8=3.2\times10^{11}$字节的内存。我们可以利用 Web 图的一些特性将上述内存的需求压缩到 10%以下。假定每个网页都用

唯一的整数表示，首先建立一个类似于倒排索引的邻接表，其每行都对应一个网页，并按照其对应的整数大小来排序。任一网页 p 对应的行中包含的也是一系列整数的排序结果，每个整数对应的是链向 p 的网页编号。这张邻接表允许应答类似于“哪些网页指向 p”的查询。以同样的方法，可以建立所有指向 p 网页的邻接表。

原始的表示方法中，均采用 32 比特整数位来表示每个链接的源页面和目标页面，而上述这种邻接表的表示方法能够将原始表示的空间降低 50%。新的方法是从网页中链出的链接来组成邻接表，此技术容易应用到链接网页的邻接表上。为了进一步减少上表的存储空间，可以采用以下几种思路。

（1）表中的相似度：表格中很多行的公共相似元素。因此，如果将多个相似行表示成一个原型，那么其他相似行就可以采用这个原型来简洁表示。

（2）局部性：某个网页会链接到其相邻的网页，比如链接到同一主机的网页。这意味着，如果对链接目标进行编码时，往往可以通过使用小数点来达到节省空间的目的。

（3）在排序表中使用间隔编码：不直接存储链接目标的编号，而是存储与其前一个元素的偏移。

10.3.8 Web 图

可以将整个静态 Web 看成是静态 HTML 网页通过超链接互相连接而成的有向图，其中每个网页是图的顶点，而每个超链接则代表一个有向边。

图 10-10 为两个顶点通过链接构成的 Web 图，每个顶点代表一个网页，A 网页上有一个超链接指向 B。将所有这样的顶点和有向边集合称为 Web 图。图 10-10 还表明，在 A 网页上的超链接周围还有一些文本，大部分网页链接的实际情况也是如此。这些文本通常被嵌在〈a〉标签（称为锚）的 href 属性中。该有向图也有可能不是一个强连通(strongly connected)图，也就说，从一个网页出发，沿着超链接前进，有可能永远不会到达另外某个网页。将指向某个网页的链接称为入链接(in-link)，而从某个网页指出去的链接称为出链接(out-link)。一个网页的链入数目被称为这个网页的入度(in-degree)，在一系列研究中得到的网页的平均入度从 8～15 不等。同样，我们可以定义某个网页的出链接数目为其出度(out-degree)。图 10-11 给出了展示这些概念的一个例子。

链接分析的研究主要基于两个基本思考点：①指向页面 B 的锚文本是对 B 的一个很好的描述；②A 到 B 的超链接表示 A 的作者对 B 的认可。当然，并非所有情况都会如此，比如，某个网站网页中的很多链接源于通用模板的使用。例如，大部分公司网站的每个网页都有一个链接指向版权声明页面。这种链接显然不代表认可的意义。因此，链接



图 10-10　两个顶点通过链接构成 Web 图

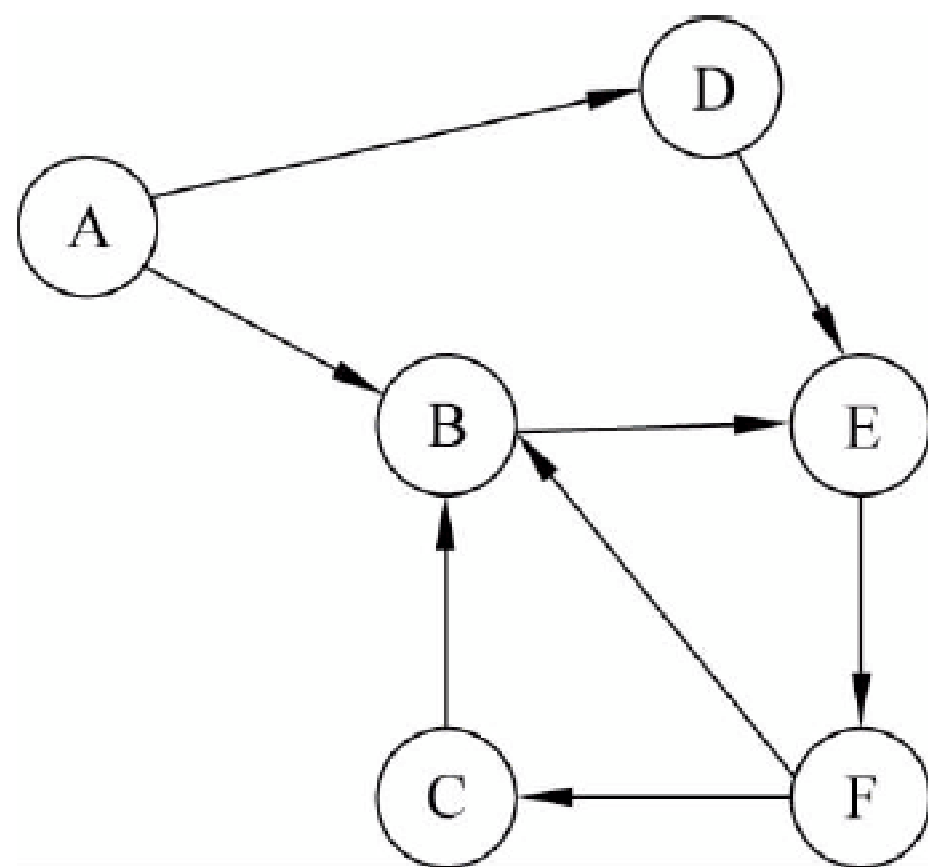


图 10-11　一个小型 Web 图例子

分析算法在实施过程中通常会去掉这些“内部”的链接。锚文本和 Web 图下面来自某个网页的 HTML 代码片段给出了一个指向期刊 Journal of the ACM 的链接：Journal of the ACM。这个例子中，链接指向页面 www.acm.org/jacm/，其锚文本为 Journal of the ACM。显然，在这个例子中锚文本是对目标页面的文字描述，但是目标 B=http://www.acm.org/jacm/本身除了其他有关期刊的信息外也包含了这段文字描述。

那么，锚文本到底起什么作用呢？Web 上随处可见的一个现象是很多网页(如图 10-11 所示的目标网页 B)的内容并不包含对自身的精确描述。很多情况下，问题主要是出在网页的设计者对网页内容的选择上。这个问题对于公司网页来说更加普遍，因为它们往往是用做商业宣传而不是介绍公司内容。尽管 IBM 被普遍认为是世界上最大的计算机制造商，但是其公司 www.ibm.com 的 HTML 代码的任何地方都不包含词项 computer。类似地，Yahoo！主页 www.yahoo.com 的 HTML 代码中也不包含单词 portal。

因此，Web 网页本身携带的词项和用户用于描述同一网页的词项之间往往存在着一定的差异。因此，Web 搜索者不一定要使用网页中的词项来对网页进行查询。另外，很多 Web 网页中的图形和图像十分丰富，并且(或者)在图像中嵌入了文字。这种情况下，采集时进行的 HTML 分析就无法抽出文本来构建网页索引，则解决方法是用锚文本来取代，通过它就可以聚集多个 Web 网页创建者的集体力量。

很多指向 www.ibm.com 链接上的锚文本都包含单词 computer，这个事实就可以为 Web 搜索引擎所使用。比如，锚文本中的词项就可以作为索引目标网页的词项。因此，词项 computer 的倒排记录表中就会包含文档 www.ibm.com，而词项 portal 的倒排记录表也同样会包含文档 www.yahoo.com。这时通过一个特别的指示器来表示这些词项出

现在锚文本中而不是页内文本中。同页内词项一样，通常也会基于词频来计算锚文本词项的权重。

锚文本的使用会产生一些有趣的副作用，例如，在大部分 Web 搜索引擎中用“big blue”来搜索时，IBM 公司的主页都会出现在排名靠前的结果中，因为这与很多人提到 IBM 时所常常采用的绰号是一致的。另外，网上已有并会持续存在这样的实例：当用类似“evil empire”的词项在 Web 搜索引擎中搜索时，这些贬义的锚文本往往会导致意料之外的结果。这种现象能够在针对某些特定网站进行的精心策划活动中得到。这种刻意策划的锚文本可能是一种作弊形式，某个网站可以通过构造具有误导性的锚文本来指向自己，从而提高在某些查询词项上的排名。检测这些对锚文本的滥用是 Web 搜索引擎所从事的另外一种必要工作。锚文本周围窗口中的文本（有时被称为扩充的锚文本，extended anchor text）常常也可以当成锚文本的扩充来等同使用。

10.4　主要网页排序算法

网页排序算法是搜索引擎的一个核心支撑技术，目的是依据一定的网页内容关系（例如链入、链出、共同链等）或网页用户行为（用户点击量、浏览时间、下载次数、页面的用户评论数等），与用户的网页检索需求进行匹配，并依据排序规则对相关网页进行合理排序，把排序后的检索结果反馈给用户。

网页排序算法（PageRank）最早是由斯坦福大学的博士研究生 Sergey Brin 和 Lawrence Page 首次提出的一种算法，它对网页质量进行评价，为每个网页赋予一个衡量其重要性的权值（PR 值），并最后应用于检索结果的排序。PageRank 的基本思想来源于传统文献计量学中的文献引文分析方法。

传统的文献引文分析认为，一篇学术论文的重要性及质量可以通过其他学术论文对其进行引用的数量来衡量，即被其他学术论文引用得越多，则这篇文章就显得越重要。PageRank 应用传统的文献引文分析思想，提出了一个假设，即网页的重要性和质量可以通过其他网页对其超文本链接的数量来衡量。具体来说，假如网页 A 有一个指向网页 B 的链接，则意味着网页 A 认为网页 B 是重要的。假如有 10 个网页指向 A 网页，而指向网页 B 的链接却只有 2 个，则说明网页 A 比网页 B 更加重要。

在计算网站排名时，PageRank 会将网站的外部链接数考虑进去。可以认为：一个网站的外部链接数越多，其 PR 值就越高；外部链接站点的级别越高（假如 Macromedia 的网站链到你的网站上），网站的 PR 值就越高。例如，ABC. COM 网站上有一个 XYZ. COM

网站的链接，那么 ABC.COM 网站必须提供一些较好的网站内容，从而 Google 会把来自 XYZ.COM 的链接作为它对 ABC.COM 网站投的一票。你可以下载和安装 Google 的工具条来检查你的网站级别(PR 值)。

10.4.1　PageRank 网页排序算法

PageRank 算法的基本思想来源于"从许多优质网页链接过来的网页，必定还是优质网页"这一回归关系。网页 A 链接到网页 B，就认为网页 A 为网页 B 投了一票。

PageRank 最初的算法描述如下：网页 A 的 PageRank 值为

$$PR(A) = (1-d) + d(PR(T_1)/PC(T_1) + \cdots + PR(T_n)/PC(T_n)) \qquad (10\text{-}1)$$

其中，d 为阻尼系数，且 $0<d<1$；$T_1,T_2,\cdots,T_n$ 表示链接到 A 的所有 n 个网页；$PR(T_1)$表示 T_1 的 PageRank 值；$PC(T_1)$表示 T_1 页面上的总链接数。而用户单击页面上链接的概率，则由页面上的链接数确定，即式(10-1)中的 $PR(T_1)/PC(T_1)$，阻尼系数 d 的引入是为了降低这一概率，因为用户不可能无限制地单击链接，常常会随机转到其他页面。

10.4.2　Topic-Sensitive PageRank 算法

Topic-Sensitive PageRank 算法是 PageRank 的一个相关算法。由于 Internet 上的内容千差万别，涵盖众多不同的领域和主题。同样一个查询主题词如"汽车"，可能用户 U_1 是想买一台汽车，他感兴趣的是汽车品牌、价格；而用户 U_2 是想参加与汽车相关的运动，他感兴趣的是与汽车相关的运动项目和赛事。因此要想给用户返回更为准确的查询信息就有必要基于不同的主题来对页面排序。最初的 PageRank 算法中没有考虑主题相关的因素参与排序。主题敏感 PageRank 算法(topic-sensitive PageRank，TSPR)正是在这种背景下提出来的。

TSPR 核心思想就是通过离线计算，计算出一个 PageRank 向量集合(在 PageRank 算法中，仅计算一个 PageRank 向量)，该集合中的每一个向量与某一主题相关，即计算某个页面关于不同主题的得分。例如某个网页在教育主题方面的得分为 a，在体育主题方面的得分为 b……

具体来说，TSPR 也可分为两个主要阶段。

(1) 主题相关的 PageRank 向量集合的计算。先将所有页面的内容划分为 16 个主题，根据 Crawler 搜集来的网页计算该网页在不同主题的得分情况，即不同的 PageRank 向量。

(2) 在线查询,主题确定。在线查询阶段,先根据用户的搜索请求确定用户的Context(用 q' 表示);然后使用式(10-2)计算用户的 Context 属于不同主题 c_j 的概率 $P(c_j|q')$;最后使用式(10-3)计算网页的综合得分 S_{qd},并根据该得分进行页面排序。式(10-3)中的 rank_{jd} 即页面 d 在主题 c_j 的得分情况。

$$P(c_j \mid q') = P(c_j) \times P(q' \mid c_j)/P(q') \propto P(c_j) \times \prod_i P(q_i' \mid c_j) \tag{10-2}$$

$$S_{qd} = \sum_j P(c_j \mid q') \times \text{rank}_{jd} \tag{10-3}$$

根据用户的查询请求和相关 Context 判断用户查询相关的主题(即用户的兴趣取向),从而提高返回结果的准确性无疑是一种有效的方法。

遗憾的是,TSPR 并没有利用主题的相关性来提高链接得分的准确性。事实上对于网页类别的划分可以更有效地计算链接的价值和权威性。例如评阅论文时,经常需要填写对相关领域的熟悉程度。也就是说,评阅者对论文所属的领域越熟悉,则评阅者所给出的评分越可信,从而在最后的计算中拥有更高的权重。

对于网页之间的链接分析与上述论文评阅的例子类似。可以把网页 A 指向网页 B 的链接视为 A 对 B 的评分;若 A 与 B 的内容是相近的,则 A 的评分更为可信。例如一个教育相关的网站 A 指向另一个教育相关的网站 B,则比一个娱乐相关的网站 C 指向教育相关的网站 B 更为权威、可信。

10.4.3 Hilltop 算法

Hilltop 算法的指导思想与 PageRank 是一致的,即通过链接的数量和质量来确定搜索结果的排序权重。与 PageRank 不同的是,在 Hilltop 中仅考虑那些专家页面(expert sources),即专门用于引导人们浏览资源的页面。Hilltop 在收到一个查询请求时,首先根据查询的主题计算出一列相关性最强的专家页面,然后根据指向目标页面的非从属专家页面的数量和相关性来对目标页面进行排序。目标页面的排序得分反映了与查询主题相关的最好的独立专家页面的集体意见。若在此过程中,Hilltop 无法得到一个足够大的专家页面集合,则返回空值。Hilltop 算法主要包含两个步骤。

(1) 专家页面搜索。所谓专家页面,就是关于某个主题的包含着很多非从属页面链接的网页。非从属页面是指两个页面分别属于两个来自非从属组织的作者。在预处理阶段,由搜索引擎的 Crawler 搜集来的网页的一个子集被辨识为专家页面集。

辨识专家页面的关键主要有:①剔除从属页面;②选择专家页面(Out-Link 大于阈值 k);③对专家页面进行索引。

当收到一个查询时，从专家页面集中挑选出与查询主题相关的专家页面子集。

(2) 目标页面排序。Hilltop 算法认为“一个目标页面在某个查询主题是权威的，当且仅当有一些与该查询主题相关的最好的专家页面指向该目标页面”。

然而，Hilltop 在应用中还存在如下一些问题。

专家页面的搜索和确定对算法起关键作用，专家页面的质量决定了算法的准确性；而专家页面的质量和公平性在一定程度上难以保证。同时 Hilltop 忽略了大多数非专家页面的影响。在 Hilltop 的原型系统中，专家页面只占到整个页面的 1.79%(2.5～140M)，在一定程度上并不能很好地反映整个 Internet 的民意。

Hilltop 算法在无法得到足够的专家页面子集时（小于两个专家页面），返回为空，即 Hilltop 适合于对查询排序进行求精，而不能覆盖。这意味着 Hilltop 可以与某个页面排序算法结合，提高精度，而不适合作为一个独立的页面排序算法。Hilltop 中根据查询主题从专家页面集合中选取与主题相关的子集也是在线运行的，这与前面提到的 HITS 算法一样会影响查询响应时间。随着专家页面集合的增大，算法的可伸缩性存在不足之处。

10.4.4　HITS 算法

HITS 算法是在 20 世纪 90 年代末提出的一种链接分析算法，它将网页的质量评估结果反映在对每个网页给出的两个评价数值——内容权威度（authority）和链接权威度（hub）上。内容权威度与网页自身提供的内容质量相关，被越多网页所引用的网页，其内容权威度越高；相对应地，链接权威度与网页提供的超链接的质量相关，引用内容质量高的越多的网页，其链接权威度越高。

HITS 算法的具体实现是一个“迭代→收敛”过程。HITS(hyperlink-induced topic search)算法与 PageRank 算法是同期由康奈尔大学的 Kleinberg 提出的，它是一种基于 Web 结构挖掘的算法。算法认为网页页面有两个方面的属性：一个是权威性（authority），被其他网页指向的属性，用 $A(T)$ 表示；另一个是中心性（hub），指向其他网页的属性，用 $H(T)$ 表示。权威性 $A(T)$ 用指向自己的网页 T_a 的中心性 $H(T_a)$ 衡量，中心性 $H(T)$ 用自己指向的网页 T_b 的权威性 $A(T_b)$ 衡量，a、b 为自然数。如下：

$$A(T) = \sum_{a=1}^{m} H(T_a) \tag{10-4}$$

$$H(T) = \sum_{b=1}^{n} A(T_b) \tag{10-5}$$

其中，m、n 分别为对应的网页数量。由公式可以得出，权威性和中心性是相互作用

的,高权威性网页是由很多高中心性网页所链接的,同时高中心性网页也必然链向很多高权威性网页。用户查询过程中,系统首先根据输入的关键词得到最相关的一组网页集合形成根集,再对其进行上下扩展,增加它所链接的和链向它的网页地址。然后通过根集特征与扩展集特征的对比,完成对扩展集内网页的筛选,去掉不相关和差别较大的网页。最后计算扩展集内网页的权威值和中心值,并依据此值进行排序。

10.4.5 SALSA 算法

PageRank 算法是基于用户随机地向前浏览网页的直觉知识,HITS 算法考虑的是 Authority 网页和 Hub 网页之间的加强关系。实际应用中,用户大多数情况下是向前浏览网页,但是很多时候也会回退浏览网页。基于上述直觉知识,R. Lempel 和 S. Moran 提出了 SALSA(stochastic approach for link-structure analysis)算法。该算法考虑了用户回退浏览网页的情况,保留了 PageRank 的随机漫游和 HITS 中把网页分为 Authority 和 Hub 的思想,取消了 Authority 与 Hub 之间的相互加权关系。

具体算法如下:

(1) 与 HITS 算法的第一步一样,得到根集并且扩展为网页集合 T,并除去孤立节点。

(2) 从集合 T 构造无向图 $G'=(V_h,V_a,E)$:

$$V_h = \{S_h \mid S \in C_{\text{and out-degree}}(S) > 0\}]\!](G' \text{ 的 Hub 边}) \tag{10-6}$$

$$V_a = \{S_a \mid S \in C_{\text{and out-degree}}(S) > 0\}(G' \text{ 的 Authority 边}) \tag{10-7}$$

$$E = \{(S_h, r_a) \mid S > r \text{ in } T\} \tag{10-8}$$

这就定义了两条链:Authority 链和 Hub 链。

(3) 定义两条马尔可夫链的变化矩阵,也就是随机矩阵,分别是 Hub 矩阵 $\boldsymbol{H}$ 和 Authority 矩阵 $\boldsymbol{A}$。

$$H_{i,j} = \sum_{K,K \in F(i) \cap F(j)} (1/\mid F(i) \mid) \times 1/B(K) \tag{10-9}$$

$$A_{i,j} = \sum_{K,K \in B(i) \cap B(j)} (1/\mid B(i) \mid) \times 1/F(K) \tag{10-10}$$

(4) 求出矩阵 $\boldsymbol{H}$ 和 $\boldsymbol{A}$ 的主特征向量,得到对应马尔可夫链的静态分布。

(5) $\boldsymbol{A}$ 中值大者对应的网页就是所要找的重要网页。

SALSA 算法没有 HITS 中相互加权的迭代过程,计算量远小于 HITS。SALSA 算法只考虑直接相邻的网页对自身 AH 的影响;而 HITS 是计算整个网页集合 T 对自身 AH 的影响。

试验结果表明，HITS 算法结果集中于主题的某个方面。而 SALSA 算法的结果覆盖了多个方面，也就是说，对于 TKC 现象，SALSA 算法比 HITS 算法有更高的健壮性。

10.4.6 BFS 算法

SALSA 算法计算网页的 Authority 值时，只考虑网页在直接相邻网页集中受欢迎程度，忽略了其他网页对它的影响。HITS 算法考虑的是整个图的结构，特别地经过 n 步以后，网页 i 的 Authority 的权重是 $|\mathrm{BF}n(i)|/|\mathrm{BF}n|$。$\mathrm{BF}n(i)$ 为离开网页 i 的 $(\mathrm{BF})n$ 的路径数目，即网页 $j<>i$，对 i 的权值贡献等于从 i 到 j 的 $(\mathrm{BF})n$ 路径数量。如果从 i 到 j 包含有一个回路，那么 j 对 i 的贡献将会呈指数级增加，这并不是算法所希望的，因为回路可能不是与查询相关的。

Allan Borodin 等人提出了 BFS(backward forward step)算法，它既是 SALSA 的扩展情况，也是 HITS 的限制情况。其基本思想是，SALSA 只考虑直接相邻网页的影响，BFS 扩展到考虑路径长度为 n 的相邻网页的影响。在 BFS 中，$\mathrm{BF}n(i)$ 被指定表示能通过 $(\mathrm{BF})n$ 路径到达 i 的节点集合，这样 j 对 i 的贡献就依赖于 j 到 i 的距离。BFS 采用指数级降低权值的方式，节点 i 的权值计算如下：

$$a_i = 2^{n-1} \mid B(i) \mid + 2^{n-1} \mid \mathrm{BF}(i) \mid + 2^{n-3} \mid \mathrm{BFB}(i) \mid + \cdots + \mid \mathrm{BFB}^n(i) \mid \quad (10\text{-}11)$$

算法从节点 i 开始，第一步向后访问，然后继续向前或向后访问邻居；每一步遇到新的节点加入权值计算，节点只有在第一次被访问时加入进去计算。

10.4.7 PHITS 算法

D. Cohn and H. Chang 提出了计算 Hub 和 Authority 的概率统计法 PHITS (probabilistic analogue of the HITS)。在这个模型中一个潜在的因子或主题 z 影响了文档 d 到 c 的一个链接。PHITS 算法进一步假定，给定因子 z，文档 c 的条件分布 $P(c|z)$ 存在，并且给定文档 d，因子 z 的条件分布 $P(z|d)$ 也存在。

$$P(d,c) = P(d) \times P(c \mid d) \quad (10\text{-}12)$$

其中，

$$P(c \mid d) = \sum_z P(c \mid z) \times P(z \mid d) \quad (10\text{-}13)$$

根据这些条件分布，提出了一个可能性函数 L：$L = \prod_{(d,c) \in \boldsymbol{M}} P(d,c)$。$\boldsymbol{M}$ 是对应的连接矩阵。

PHITS 算法使用 Dempster 等人提出的 EM 算法分配未知的条件概率，使得 L 最大

化,即最好地解释了网页之间的链接关系。算法要求因子 z 的数目事先给定。

本章小结

Web 是 Internet 最基本、最广泛的应用服务,也是最主要的信息资源类型。对于信息用户而言,直接面对的 Web 信息获取工具就是网络搜索引擎,Google、Baidu 等搜索引擎是 Web 信息采集与搜索技术的典型代表。搜索引擎(search engine)是指根据一定的策略、运用特定的计算机程序搜集互联网上的信息,在对信息进行组织和处理后,为用户提供检索服务的系统。

一般情况下将搜索引擎分为采集器、索引器、检索器和用户接口四个部分。通常搜索引擎有目录搜索引擎、全文搜索引擎、元搜索引擎、集合式搜索引擎、垂直搜索引擎。而智能搜索引擎是在传统搜索引擎基本结构的基础上,增加了相关技术或者相关系统来优化整个搜索引擎的综合检索系统,包括基于本体的智能搜索引擎、基于知识库系统的智能搜索引擎、基于语义关联的智能搜索引擎等类型。

搜索引擎技术原理的种类较多,主要因其应用的信息采集算法原理和索引技术的不同而不同。目前,搜索引擎的主要支撑技术有分词技术、网络蜘蛛、索引技术、词频相关指数、主动推理、本体知识系统、专家系统等类型。

Web 信息采集技术被广泛应用于搜索引擎检索、站点结构分析、页面有效性分析、Web 图进化、内容安全检测、用户兴趣挖掘以及个性化信息获取等多种服务和研究当中。Web 采集是从 Web 中收集网页的过程,这些网页用于索引从而为搜索引擎提供支持。采集的目标是尽可能高效地采集更多数目的有用页面,并同时获得连接这些页面的链接结构。

Web 采集器架构主要由五类模块构成:待采集 URL 池、DNS 解析模块、抓取(fetch)模块、分析(parse)模块、URL 去重模块。分布式信息检索是指由检索代理程序将检索任务同时提交给网络上的多个主机,由位于这些主机上的检索程序分别独立检索并将检索结果返回到检索代理程序,经过整理后显示给用户。

PageRank 最早是由斯坦福大学的博士研究生 Sergey Brin 和 Lawrence Page 首次提出的一种算法,它对网页进行评价,为每个网页赋予一个衡量其重要性的值,并最后应用于检索结果的排序。PageRank 的基本思想主要来自传统的文献计量学中的文献引文分析。主要的 PageRank 排序算法有 PageRank 网页排序算法、Topic-Sensitive PageRank 算法、Hilltop 算法、HITS 算法、SALSA 算法、BFS 算法、PHITS 算法等排序算法。

本章思考与练习题

1. 举例说明你是如何使用你所熟悉的一个网络搜索引擎的。
2. 搜索引擎由哪几个部分组成?
3. 通常搜索引擎分为哪几种类型?
4. 说明元搜索引擎与垂直搜索引擎的含义。
5. 基于本体的智能搜索引擎的主要结构模块有几个部分?
6. 用图示说明基于知识库系统的智能搜索引擎的结构原理。
7. 说明基于语义关联的智能搜索引擎的功能模块。
8. 说明基于语义关联的智能搜索引擎的工作步骤。
9. 搜索引擎有哪些主要支撑技术?
10. 说明分词技术的含义。
11. 什么是网络蜘蛛?
12. 索引技术的作用与含义有哪些?
13. 词频相关指数的含义有哪些?
14. 自动推理技术的含义有哪些?
15. 举例说明本体知识系统的含义。
16. 专家系统的内涵有哪些?
17. 如何理解 Web 信息采集的含义?
18. Web 信息采集器应该提供哪些功能?
19. Web 采集器架构主要由哪五类模块构成?
20. 简述 URL 的采集流程。
21. 递归查询的含义有哪些?
22. 举例说明分布式信息检索的含义。
23. 说明 Web 图的作用与意义。
24. 主要的 PageRank 排序算法有哪些? 简述各自的基本原理。

第三部分

信息检索素养实践应用篇

对于大学生尤其是研究生而言，信息检索素养最直接的体现就是服务于他们的自主学习、协同学习、研究性学习、探究与发现性学习等主动性与高层次特征的学习活动及其学习过程。本书第三部分"信息检索素养实践应用篇"正是基于这一目的进行教学设计与内容编著，内容包括第 11 章、第 12 章和第 13 章，其中包含了大量丰富的图例与实例阐述，以便于学习者结合自身的信息需求实际，理论联系实际，举一反三。第三部分"信息检索素养实践应用篇"作为理论教学内容的同时，可以同本课程实验与实践教学要求相结合，把信息检索素养教育融入理论与实践相互贯通的教学实践过程中。

互联网是一个海量的信息世界，各类信息资源十分丰富，如何快速准确地在网络上检索并获取所需信息，实现用户网络信息需求的满足，对于网络化时代的每一位网民而言，都是一个非常重要的问题。作为新时代的大学生，应用搜索引擎去充分发现、认识、查询、获取和有效利用网络信息，不仅是大学生信息检索素养的重要组成部分，也是开展自主学习、协同学习、探究性与研究性学习的基础性信息素养及其内在要求。作为大学生信息检索素养能有落地生根的一个重要基础，第 11 章"常用搜索引擎的检索应用"详细阐述和说明了百度、搜狗、Google、Infoseek 和 Yahoo!的各种检索应用，尤其是这些引擎的高级检索(或专业检索)与应用，从而实现信息用户对信息查全率和查准率的更高

要求。

对于大学生或科技工作者而言，特种信息资源是指出版发行和获取途径都比较特殊的科技类信息资源，通常也指的是除了普通图书信息资源和期刊学术论文信息资源之外的特种科技信息资源。第12章“特种信息资源检索”详细阐述了中外重要的会议文献信息资源、科技报告信息资源、专利信息资源、学位论文信息资源、标准信息资源、科技档案信息资源、政府出版物信息资源七大类特种信息资源。特种信息资源特色鲜明、内容广泛、数量庞大、学习与研究及其参考价值高，在整个信息资源检索及其利用过程中起着非常重要的作用。

图书与学术期刊论文信息资源是大学生最主要的信息检索与利用对象，也是各个高校图书馆投入资金比例最大、收藏量最丰富、占用馆藏最多、连续性购买强度最高并提供基础性服务支持与保障最有力的主要资源。第13章“图书与学术期刊论文信息资源检索”着重阐述了主要和典型中外图书与学术期刊论文信息资源检索与应用。主要和典型中外图书与学术期刊论文信息资源包括中国国家图书馆联机公共目录查询系统、CALIS联合目录公共检索系统、北京大学图书馆公共查询系统、清华大学图书馆馆藏目录检索系统、典型中文数字图书检索——超星数字图书馆、CNKI中国学术期刊网检索、维普中文科技期刊数据库检索、CADAL外文图书检索、World eBook Library检索、ebrary(电子图书馆)检索、OCLC FirstSearch检索、Web of Science数据库检索、IEL数据库检索、EBSCO学术资源平台检索、Wiley在线图书馆检索等。

第 11 章　常用搜索引擎的检索应用

互联网是一个海量的信息世界，各类信息资源十分丰富，如何快速准确地在网络上检索并获取所需信息，实现用户信息需求的满足，对于网络化时代的每一位网民而言，都是一个非常重要的问题。作为新时代的大学生，应用搜索引擎去充分发现、认识、查询、获取和有效利用网络信息，不仅是大学生信息检索素养的重要组成部分，也是开展自主学习、协同学习、探究性与研究性学习的基础性信息素养及其内在要求。

搜索引擎（search engine）是一种网络化信息检索系统与检索应用工具，能帮助用户在浩瀚的网络资源环境中快速而高效地查询到所需要的信息。搜索引擎是一种能够通过网络接收用户的查询指令，并向用户提供符合其查询要求的信息资源网址或资源路径的智能系统。在很多搜索引擎中，利用在层次结构中的不同的高速缓存来存储的一些数据块，这是非常有用的解决频繁查询的方法。

作为普通用户而言，经常接触到的是网络搜索引擎的用户检索交互界面。用户检索交互界面是搜索引擎各种检索实现功能在用户接口层面的直接而形象的表达，屏蔽了搜索引擎所应用的各种检索原理、检索技术与数学逻辑过程。用户检索交互界面的作用是接收用户的查询输入、显示查询结果、提供相关反馈信息。用户检索界面包括简单检索界面和高级检索界面两类。简单检索界面只提供用户输入查询字符串的文本框，高级检索界面提供用户按照各类检索模型的查询机制，常用的检索模型有集合论模型、代数模型、概率模型和混合模型等，具体体现为逻辑运算（与、或、非等）、相近关系（相邻、近似等）、域名范围（如. edu、.com 等）、位置限定（如标题、内容等）、时间限定或信息的语种限制等。

11.1　百度搜索引擎的检索应用

1. 百度简述

百度这一公司名称便来自宋词“众里寻他千百度”。百度公司会议室名为“青玉案”，即是这首词的词牌。而“熊掌”图标来源于“猎人巡迹熊爪”的刺激，与李彦宏博士的“分析搜索技术”非常相似，从而构成百度的搜索概念，也最终成为了百度的公司图标。由于在

搜索引擎领域大都有动物形象,如 Sohu 的狐、Google 的狗,而百度也便顺理成章称为熊,百度熊也便成了百度公司的形象物。图 11-1 为百度引擎 Logo。

百度搜索引擎是目前规模最大、影响力最大、最受中文用户欢迎的中文搜索引擎。1999 年年底,百度成立于美国硅谷,它的创建者是在美国硅谷有多年成功经验的李彦宏先生,2000 年百度公司回国发展,百度的起名,来自“众里寻他千百度”的灵感,寄托着百度公司对自身技术与发展前景的信心,蕴含了“用户第一”并提供高质量网络信息搜索服务的价值追求。



图 11-1　百度引擎 Logo

2. 百度核心技术

百度搜索引擎由四个核心部分组成:蜘蛛程序、监控程序、索引数据库和检索程序。百度门户网站只需将用户查询内容和一些相关参数传递到百度搜索引擎服务器上,后台程序就会自动工作并将最终结果返回给网站。百度搜索引擎使用了高性能的“网络蜘蛛”程序自动地在互联网中搜索信息,可定制、高扩展性的调度算法使得搜索器能在极短的时间内收集到最大数量的互联网信息。

百度搜索引擎采用了先进的“链接分析”(link analysis)技术,这种技术将传统信息学中的引文索引技术同 Web 中最基本的“超级链接分析”技术相结合,在查找的准确性、查全率、更新时间、响应时间等方面与其他技术相比都有明显优势。

同时,百度应用内容相关度评价技术,并且运用了中文智能语言处理方法,依靠字与词的不同切割方法,弥补了单纯依靠字或词的引擎技术的固有缺陷,并且能够在不同的编码之间转换,这就使得简体字和繁体字的检索结果自然结合,相得益彰。

3. 百度引擎信息服务产品

1) 最新上线的信息服务

截止到 2016 年 4 月,最新上线的信息服务包括度秘、宝宝知道、百度优课、百度春华 APP 推广、百度 MALL 7 种信息服务产品,见图 11-2。



图 11-2　截止到 2016 年 4 月百度最新上线信息服务类型

2）百度信息搜索服务

百度信息搜索服务产品丰富，包括搜索网页、视频、音乐、新闻、图片、软件等 18 种。见图 11-3。



图 11-3　百度信息搜索服务类型

3）百度导航服务

百度导航服务主要有三类产品，即 hao123、网站导航和百度口碑（评论信息搜索服务）。见图 11-4。



图 11-4　百度导航信息服务类型

4）百度社区信息服务

百度社区信息服务类型丰富，包括百度文库、百度网盘、百度知道、百度贴吧等 23 种。见图 11-5。

5）娱乐游戏信息服务

娱乐游戏信息服务产品包括 91 手游、百度游戏、百度应用、百度爱玩、百度电视游戏 5 种。见图 11-6。

6）软件工具信息服务

软件工具信息服务包括百度传课、百度输入法、百度浏览器等 10 种。见图 11-7。



图 11-5　百度社区信息服务类型



图 11-6　娱乐游戏信息服务



图 11-7　百度软件工具信息服务

7）百度移动端信息服务

百度移动端信息服务内容包括百度糯米、百度理财、手机输入法、手机助手、手机地图等 12 种。见图 11-8。

8）百度其他专题信息服务

百度其他专题信息服务包括 91 门户、苹果园、安卓网、百度公益、百度营销大学、百度认证等 7 种。见图 11-9。



图 11-8　百度移动端信息服务



图 11-9　百度其他专题信息服务

4. 百度网页搜索

1）百度快照

如果无法打开某个搜索结果，或者打开速度特别慢，该怎么办？"百度快照"能帮您解决问题。每个未被禁止搜索的网页，在百度上都会自动生成临时缓存页面，称为"百度快照"。图 11-10 是输入检索词"云计算"后的结果实例。

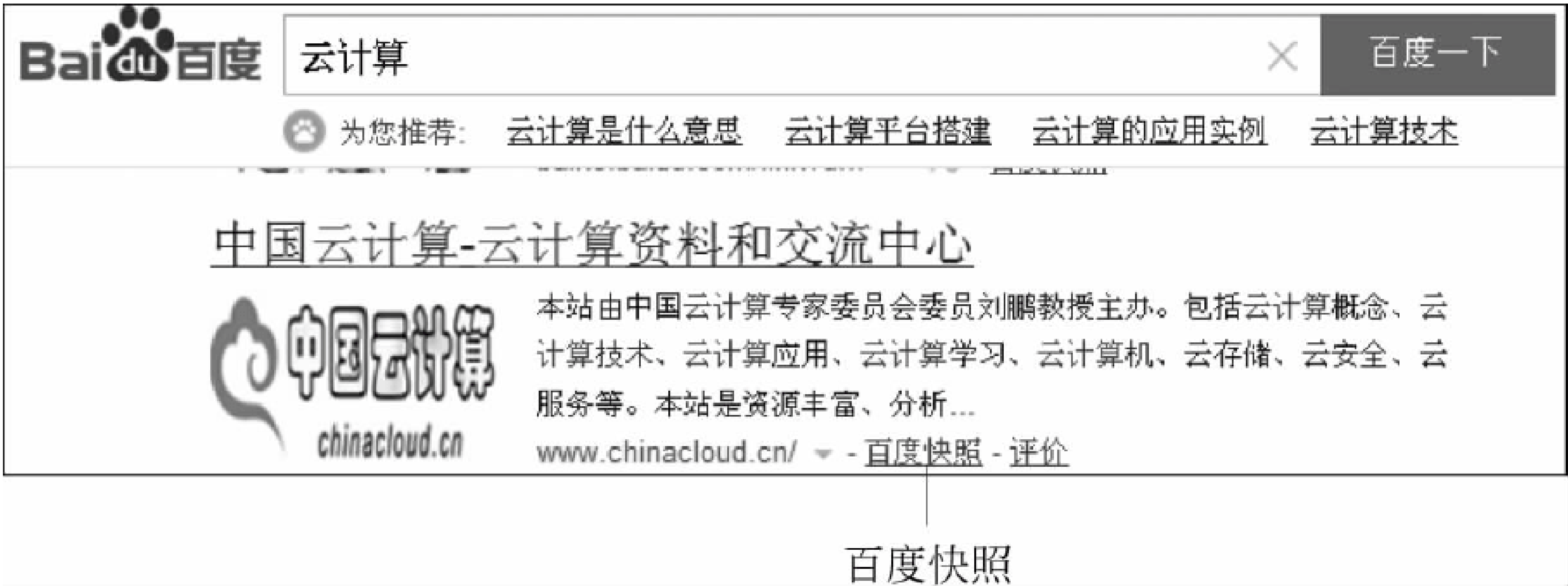


图 11-10　输入检索词"云计算"后的结果实例

当您遇到网站服务器暂时故障或网络传输堵塞时，可以通过"快照"快速浏览页面文本内容。百度快照只会临时缓存网页的文本内容，所以那些图片、音乐等非文本信息，仍是存储于原网页。

当原网页进行了修改、删除或者屏蔽后，百度搜索引擎会根据技术安排自动修改、删除或者屏蔽相应的网页快照。

2）拼音输入替代汉字

在不知道汉字的情况下，输入拼音可以吗？如果只知道某个词的发音，却不知道怎么写，或者嫌某个词拼写输入太麻烦，该怎么办？百度拼音提示能帮您解决问题。

只要输入查询词的汉语拼音，百度就能把最符合要求的对应汉字提示出来。它事实上是一个无比强大的拼音输入法，拼音提示显示在搜索结果上方。例如，输入“zhurongji”，提示如下：您要找的是不是：朱镕基，检索结果如图 11-11 所示。



图 11-11　拼音输入替代汉字的检索实例

3）相关搜索

搜索结果不佳，有时候是因为选择的查询词不是很妥当。您可以通过参考别人是怎么搜的来获得一些启发。百度的“相关搜索”，就是和您的搜索很相似的一系列查询词。

百度相关搜索排布在搜索结果页的下方，按搜索热门度排序。单击这些词，可以直接获得它们的搜索结果，图 11-12 是“无人机”的相关搜索。

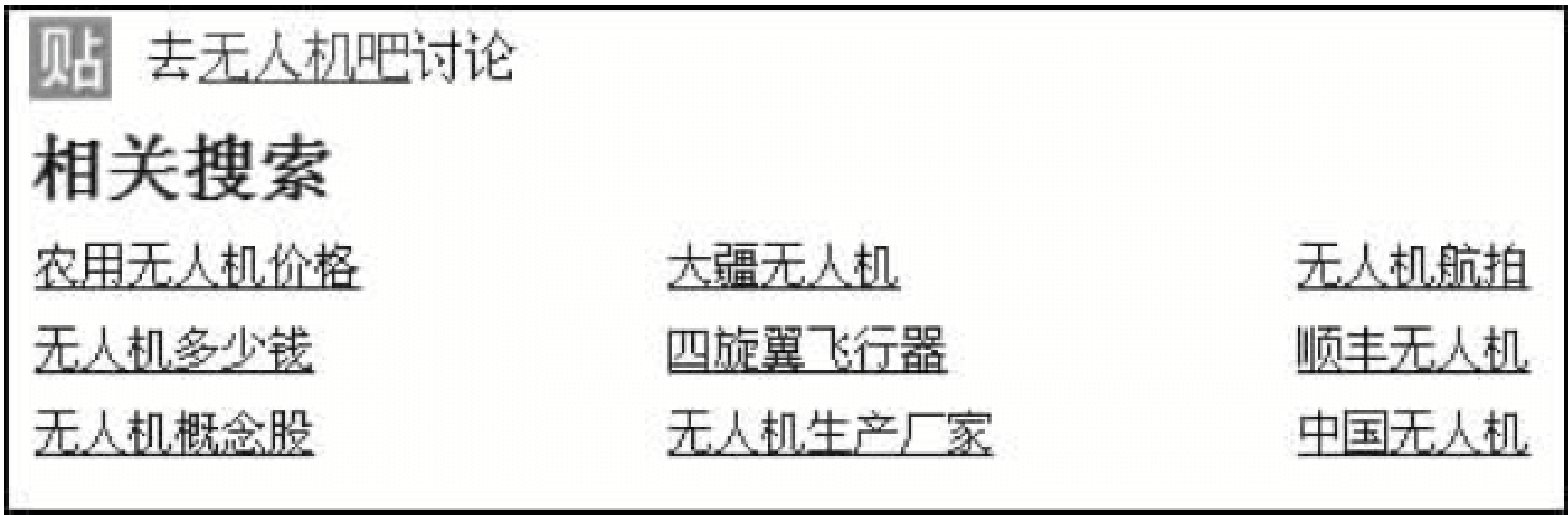


图 11-12　“无人机”的相关检索实例

4）错别字校正

由于汉字输入法的局限性，我们在检索时经常会输入一些错别字，导致搜索结果不佳。百度会给出错别字纠正提示，并且给出正常结果。错别字提示显示在搜索结果上方。

例如，输入“唐醋排骨”，提示如下：您要找的是不是：糖醋排骨，结果实例如图 11-13 所示。



图 11-13　错别字“唐醋排骨”自动校正的搜索结果实例

5）网页搜索中的英汉互译词典

百度网页搜索内嵌英汉互译词典功能。如果想查询英文单词或词组的解释，您可以在搜索框中输入想查询的“英文单词或词组”＋“是什么意思”，搜索结果第一条就是英汉词典的解释，例如，retrival 是什么意思（如图 11-14 所示）；如果您想查询某个汉字或词语的英文翻译，您可以在搜索框中输入想查询的“汉字或词语”＋“的英语”，搜索结果第一条就是汉英词典的解释，例如，龙的英语。另外也可以通过选择搜索框左下方的“百度翻译”链接 fanyi. baidu. com，到百度词典中查看想要的词典解释。



图 11-14　百度英汉互译的检索实例

6）计算器和度量衡转换

Windows 系统自带的计算器功能过于简陋，尤其是无法处理一个复杂计算式，很不方便。而百度网页搜索内嵌的计算器功能，则能快速高效地解决信息搜索过程中的计算需求。只需简单地在搜索框内输入计算式，回车即可。较为复杂计算式：log((sin(5))^2)－3＋pi 的结果如图 11-15 所示。

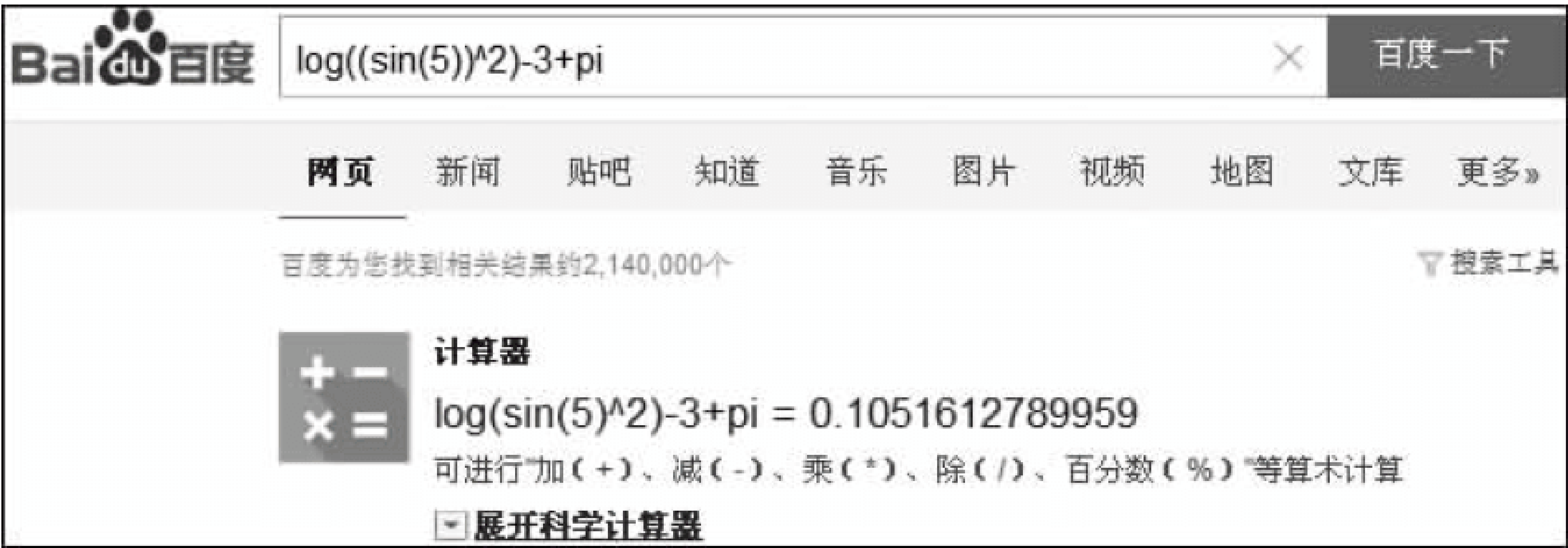


图 11-15　百度计算器的复杂计算实例图

百度计算器支持实数范围内的计算，支持的运算包括加法(＋)、減法(－)、乘法(＊或×)、除法(/)、幂运算(^)、阶乘(!)。支持的函数包括正弦、余弦、正切、对数。同时支持上述运算的混合运算。

例如，加法：3＋2，减法：3－2，乘法：3＊2，除法：3/2，阶乘：4!(4 的阶乘)，平方：4^2(4 的平方)，立方：4^3(4 的立方)，开平方：4^(1/2)(4 的平方根)，开立方：4^(1/3)(4 的立方根)，倒数：1/4(4 的倒数)，幂运算：2^8(2 的 8 次方)，常用对数：log(8)(以 10 为底 8 的对数)，以自然底数为底的对数：ln(8)(以 e 为底 8 的对数)，求弧度的正弦：sin(10)(10 弧度角正弦值)，求弧度的余弦：cos(10)(10 弧度角余弦值)，求弧度的正切：tan(10)(10 弧度角正切值)，上述运算的混合运算：log((5＋5)^2)－3＋pi，圆周率 pi＝3.141 592 65 自然底数 e＝2.718 281 83。

度量衡换算。百度支持常用的度量衡换算。方法是在搜索栏或者计算框内输入如下格式表达式：换算数量换算前单位＝？换算后单位。例如“5 公斤＝？毫克”的检索结果如图 11-16 所示。

图 11-16 中显示“度量衡换算”包括质量、长度、面积、体积、温度、压力、功率、功能/热的换算。

7）专业文档搜索

很多有价值的资料，在互联网上并非是普通的网页，而是以 Word、PDF、PowerPoint



图 11-16　百度度量转换的搜索结果实例

等格式的信息格式存在。百度支持对 Office 文档(包括 Word、Excel、PowerPoint)、Adobe PDF 文档和 RTF 文档的全文搜索。

(1) 直接搜索指定文档资料。要搜索这类文档很简单,在普通的查询词后面加一个"filetype:"文档类型限定。"Filetype:"后可以跟以下文件格式:DOC、XLS、PPT、PDF、RTF、ALL。其中,ALL 表示搜索所有这些文件类型。例如,查找经济学家樊纲先生关于收入差距方面的 DOC 资料。输入"樊纲 收入差距 filetype:doc",单击结果标题,直接下载该文档。见图 11-17。



图 11-17　直接搜索指定文档资料实例

(2) 通过文档搜索查找。可以通过百度文档搜索界面(http://file. baidu. com/),直接使用专业文档搜索功能。

(3) 查找论文网站。网上有很多收集论文的网站。先通过搜索引擎找到这些网站,然后再在这些网站上查找自己需要的资料,这是一种方案。找这类网站,简单地用"论文"做关键词进行搜索即可。例如,论文。

(4) 直接找特定论文。除了找论文网站外,也可以直接搜索某个专题的论文。一般的论文结构都有一定的规范格式,除了标题、正文、附录外,还需要有论文关键词、论文摘要等。其中,“关键词”和“摘要”是论文的特征词汇,而论文主题通常会出现在网页标题中。例如,intitle:数据挖掘,表示需要查询“数据挖掘”方面的论文信息,“数据挖掘”在论文中的关键词、摘要和标题中均出现。

(5) 百度学术搜索。如果需要搜索专业的学术论文,可以选择百度学术搜索(xuesu.baidu.com)。百度学术搜索主界面和高级搜索分别见图 11-18 和图 11-19。



图 11-18　百度学术搜索主界面



图 11-19　百度学术搜索的高级搜索

通过“百度学术”高级搜索界面设置更精准的检索词与检索项,以满足所需要的学术文档检索。包括对检索词的一些限定要求:包含全部检索词、包含精确检索词、检索词在文档中的位置(标题中、摘要中、正文中等)、作者、出版物、发布时间等约束。

8) 多个检索词组合搜索

输入多个检索词语搜索,需要在不同字词之间用一个空格隔开,可以获得更精确的搜

索结果。例如，想了解上海人民公园的相关信息，在搜索框中输入“上海 人民公园”获得的搜索效果会比输入“人民公园”的检索结果更好。见图 11-20。



图 11-20　多个检索词组合检索实例

9）善于运用“搜索框提示”

百度会根据用户的输入内容，在搜索框下方实时展示“最符合的检索提示词”。只需用鼠标单击需要的提示词，或者用键盘上下键选择想要的提示词并按回车，就会返回该检索词的查询结果。不必再费力地敲打键盘即可轻松地完成查询。

输入拼音或汉字，百度会给出最符合要求的提示。例如输入“moshou”，搜索框提示中会显示“魔兽世界”、“魔兽秘籍”等(如图 11-21 所示)；输入“kaix”，搜索框提示中会显示“开心网”、“开心农场”等；输入“百度”，搜索框提示中会显示“百度地图”、“百度空间”等。



图 11-21　百度搜索框提示的检索应用实例

默认情况下，在百度主页和搜索结果页上方的搜索框都会显示“搜索框提示”。如果用户不希望显示搜索框提示，可以在搜索框右侧“设置”列表选择“搜索设置”(如图 11-22 所示)的“搜索框提示”中选择“不显示”来关闭搜索框提示功能。关闭之后还可以在搜索框右侧设置的“搜索框提示”中选择“显示”来重新开启它。



图 11-22　百度搜索设置

显示搜索框提示时，会默认屏蔽用户浏览器的搜索框历史提示功能。如果您想恢复

浏览器的搜索框历史提示功能，请在搜索框右侧设置的“搜索框提示”中选择“不显示”（如图 11-23 所示）。



图 11-23　百度搜索框提示的个性化设置

5. 百度高级搜索和个性设置

1）高级搜索和个性设置

可以根据用户自己的检索习惯，在搜索框右侧的“设置”中，改变百度默认的搜索设定。例如搜索框提示的设置、搜索结果的每页显示数量等。百度高级搜索界面见图 11-24。



图 11-24　百度高级搜索界面

当用户在检索复杂信息需求主题并面临多个检索词时，需要确定各个检索词之间的

相互逻辑关系和检索时对每个检索词的搜索结果限定要求。

(1) 搜索结果的逻辑组配关系：包含以下的全部关键词(例如，扩大内需与收入差距的关系)，包含以下完整关键词(例如，“大学生就业政策”)，包含以下任意关键词(例如，节能环保 低碳生活)，不包括以下关键词(例如，新能源-核能)。

(2) 检索结果的时间限定：把搜索结果的网页限定在全部时间(默认)、最近一天、最近一周、最近一月或最近一年。

(3) 文档格式：所有网页与文件(默认)、. PDF、. doc、. xls、. ppt、. rtf 或所有格式的文档。

(4) 关键词的位置设定：指定查询的关键词位于网页任何地方、仅在网页的标题中或仅在网页的 URL 中。

(5) 站内搜索：限定要搜索的内容在指定的网站，例如，www. sina. com. cn。

2) 高级搜索语法运用

高级搜索语法的掌握与合理运用，对于大学生而言能够摆脱“初级傻瓜式检索”所带来的大量检索结果输出的信息筛选尴尬或评价困惑，能够提高网络信息检索的质量。

(1) intitle 语法运用。把搜索范围限定在网页标题中即 intitle 语法运用。

网页标题通常是对网页内容提纲挈领式的归纳。把查询内容范围限定在网页标题中，有时能获得良好的效果。使用的方式是把查询内容中特别关键的部分用“intitle:”连起来。

例如，找林青霞的写真，则查询式为：写真 intitle：林青霞。“intitle:”和后面的关键词之间不要有空格。

(2) site 语法运用。把搜索范围限定在特定网站中即 site 语法运用。

有时候，如果知道某个站点中有自己需要找的信息对象，就可以把搜索范围限定在这个站点中，以提高查询效率。使用的方式是在查询内容的后面加上“site:站点域名”。

例如天空网的下载软件不错，则检索式为：3D MAXs site:skycn. com。注意“site:”后面跟的站点域名不要带“http://”；另外，“site:”和站点名之间不要带空格。

(3) inurl 语法运用。把搜索范围限定在 url 链接中即 inurl 语法运用。

网页 url 中的某些信息常常有某种有价值的含义。如果对搜索结果的 url 做某种限定，就可以获得良好的查询效果。实现的方式是用“inurl:”，后跟需要在 url 中出现的查询关键词。

例如查询关于 Unity3D 游戏编程方面的信息，则查询式为：Unity3D inurl:youxi。

查询式中的“Unity3D”是可以出现在网页的任何位置的，而“youxi”则必须出现在网页 url 中。注意，“inurl:”语法和后面所跟的关键词不要有空格。检索实例如图 11-25 所示。

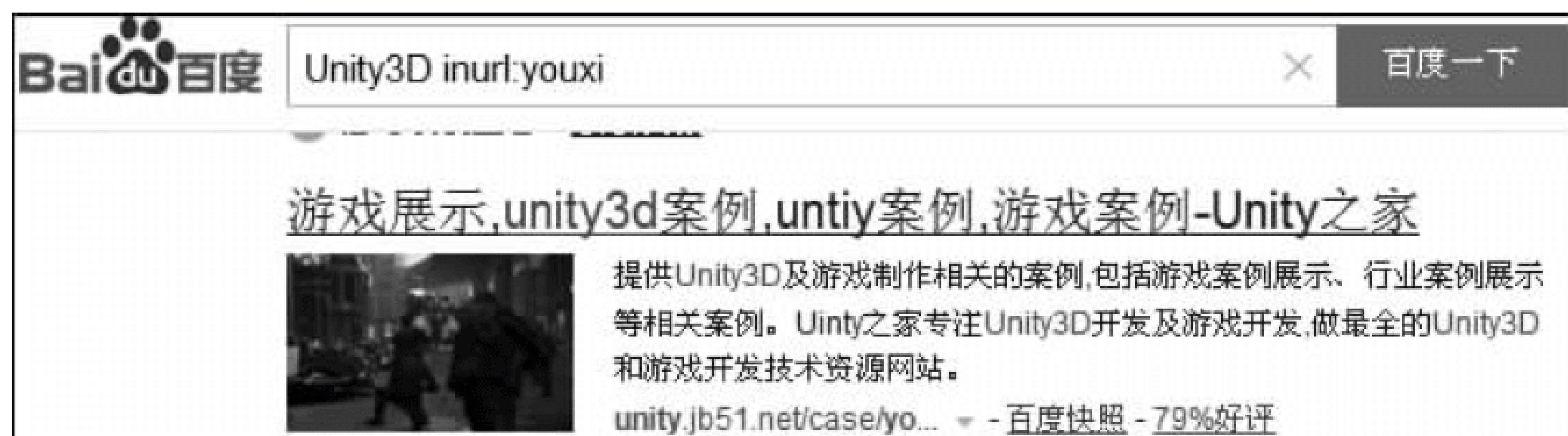


图 11-25　inurl 语法应用实例

(4) 精确匹配运算符。精确匹配符为双引号和书名号。

如果输入的查询词很长，百度在经过分析后，给出的搜索结果中的查询词可能是拆分的。如果用户对搜索结果不满意，可以尝试让百度不拆分查询词。给查询词加上双引号，就可以达到这种效果。

例如搜索中国地质博物馆，如果不加双引号，搜索结果被拆分，则返回的效果不是很理想。但是加上双引号后即“中国地质博物馆”，则获得的返回结果全部符合要求。

书名号是百度独有的一个特殊查询语法。在其他搜索引擎中，书名号会被忽略，而在百度检索过程中中文书名号是可被查询的。加上书名号的查询词有两层特殊功能：一是书名号会出现在搜索结果中，二是被书名号扩起来的内容不会被拆分。书名号在某些情况下特别有效。例如查询的名字很通俗和常用的那些电影或者小说，可能会出现歧义。比如查询电影“手机”，如果不加书名号，很多情况下出来的是通信工具的“手机”含义，而加上书名号《手机》后返回的结果就都是关于电影方面的信息。

(5) 排除语法。排除语法就是要在返回结果中去除不需要的部分，即“－”号运算符应用。

排除语法的目的是要求搜索结果中不含特定查询词。如果发现搜索结果中有某一类网页是用户不希望看见的，而且这些网页都包含特定的关键词，那么用减号语法就可以去除所有这些含有特定关键词的网页。

例如查询神雕侠侣，希望是关于电视剧方面的信息内容，却发现很多关于游戏方面的网页。那么就需要的查询式为：神雕侠侣-游戏。注意前一个关键词和减号之间必须有空格，否则减号会被当成连字符处理，而失去减号语法的搜索功能。减号和后一个关键词之间有无空格均可，检索实例如图 11-26 所示。



图 11-26　排除语法应用实例

6. 百度引擎的常用检索技巧

1）查询词的恰当选择

搜索信息最基本同时也是最有效的就是选择合适的查询词。选择查询词是一种专业知识与个人经验的积累，在一定程度上也有章可循。

（1）表述准确

百度会严格按照用户提交的查询词去搜索，因此查询词表述准确是获得良好搜索结果的必要前提。一类常见的表述不准确情况是心里想着一回事，但是搜索框里输入的检索词是另一回事。例如，要查找 2015 年国内十大新闻，查询词可以是“2015 年国内十大新闻”；但如果把查询词换成“2015 年国内十大事件”，搜索结果就不能满足原来的信息需求。

另一类典型的表述不准确是查询词中包含错别字。例如要查找林心如的写真图片，用“林心如写真”当然是没什么问题；但如果写错了字，变成“林心茹写真”，搜索结果质量就差得远了。不过百度对于用户常见的错别字输入，有纠错提示。您若输入“林心茹写真”，在搜索结果上方，会提示“您要找的是不是：林心如写真”（如图 11-27 所示）。



图 11-27　检索词表述不准确的应用实例

（2）查询词的主题关联与简练

目前的搜索引擎并不能很好地处理自然语言。因此在提交搜索请求时，用户最好把自己的想法提炼成简单的而且与希望找到的信息主题关联的查询词。例如，某三年级小

学生，想查一些关于时间的名人名言，如果查询词是“小学三年级关于时间的名人名言”。这个查询词很完整地体现了搜索者的搜索意图，但返回的查询效果并不好。绝大多数名人名言，并不规定是针对几年级学生的。因此“小学三年级”事实上和主题无关，会使得搜索引擎丢掉大量不含“小学三年级”但非常有价值的信息，而且词语“关于”也是一个与名人名言本身没有关系的词，多一个这样的词又会减少很多有价值信息；“时间的名人名言”中的“的”也不是一个必要的词，会对搜索结果产生干扰。对于检索词“名人名言”中的“名言”通常就是名人留下来的，在名言前加上名人是一种不必要的重复。因此，最好的查询词应该是“时间名言”。试着找出下述查询词的问题，并拟定更好的能满足搜索需求的查询词：所得税会计处理问题探讨、周星驰个人档案和所拍的电影。

（3）根据网页特征选择查询词

很多类型的网页都有某种相似的特征。例如，小说网页，通常都有一个目录页，小说名称一般出现在网页标题中，而页面上通常有“目录”两个字，单击页面上的链接，就进入具体的章节页，章节页的标题是小说章节名称；软件下载页，通常软件名称在网页标题中，网页正文有下载链接，并且会出现“下载”这个词等。经常搜索并且总结各类网页的特征现象，并应用查询词的选择中，就会使得搜索变得准确而高效。例如，找明星的个人资料页，一般来说明星资料页的标题通常是明星的名字，而在页面上会有“姓名”、“身高”等词语出现。比如找林青霞的个人资料，就可以用“林青霞 姓名 身高”来查询。而由于明星的名字一般在网页标题中出现，因此更精确的查询方式可以是“姓名 身高 intitle:林青霞”。intitle，表示后接的词限制在网页标题范围内。这类主题词加上特征词的查询构造方法适用于搜索具有某种共性的网页，前提是用户必须了解这种共性。

2）利用百度寻找下载软件

日常工作和娱乐需要用到大量的软件，很多软件属于共享或者自由性质，可以在网上免费下载到。百度软件中心找软件，在搜索框输入对应软件名称，例如，flashget。见图 11-28。

直接找下载页面，这是最直接的方式。软件名称加上“下载”这个特征词，通常可以很快找到下载点。例如，flashget 下载。

3）利用百度寻找问题的解决方法

我们在工作和生活中，会遇到各种各样的疑难问题。例如，计算机中毒了、被开水烫伤了等。很多问题其实都可以在网上找到解决办法。因为某类问题发生的概率是稳定的，而网络用户成千上万，于是庞大用户群中遇到同样问题的人就会很多，其中一部分人会把问题贴在网络上求助，而另一部分人可能就会把问题解决办法发布在网络上。有了



图 11-28　百度软件中心软件下载实例

搜索引擎，就可以把这些信息找出来。

找这类信息，核心问题是如何构建查询关键词。一个基本原则是：在构建关键词时尽量不要用自然语言（所谓自然语言就是我们平时说话的语言），而要从自然语言中提炼关键词。这个提炼过程并不容易，但是我们可以用一种将心比心的方式思考：如果我知道问题的解决办法，我会怎样对此做出回答。也就是说，先猜测信息的表达方式，然后根据这种表达方式取其中的特征关键词，从而达到搜索目的。

例如，我们上网时经常会遇到陷阱，浏览器默认主页被修改并锁定。这样一个问题的解决办法，我们应该怎样搜索呢？首先要确定的是，不要用自然语言。比如，有的人可能会这样搜索“我的浏览器主页被修改了，谁能帮帮我呀”，这是典型的自然语言。口语化的搜索词也可以给出适当的答案，但是这样的搜索常常得不到最想要的结果。我们来看这个问题中的核心词汇：对象是浏览器（或者 IE）的主页。事件：被修改（锁定）。“浏览器”、“主页”和“被修改”，在这类信息中出现的概率会最大，IE 可能会出现，至于锁定，用词比较专业化，不见得能出现。于是关键词中至少应该出现“浏览器”、“主页”和“被修改”，这是问题现象描述。一般情况下，只要对问题做出适当的描述，在网上基本上就可以找到解决对策。例如，浏览器主页 被修改、冲击波病毒 预防。

4）利用百度寻找英汉互译

尽管手头有英文词典，但翻词典一是麻烦速度慢，二是可能对某些词汇的解释不够详尽。中译英就更是如此了。多数词典只能对单个汉字词语做出对应的英文解释，但该解释在上下文中也许并不贴切。搜索引擎找英汉互译的一个长处就在于，可以比较上下文，

使翻译更加精确。百度本身提供了英汉互译功能，fanyi. baidu. com 提供在线翻译等的功能。图 11-29 为百度翻译实例。

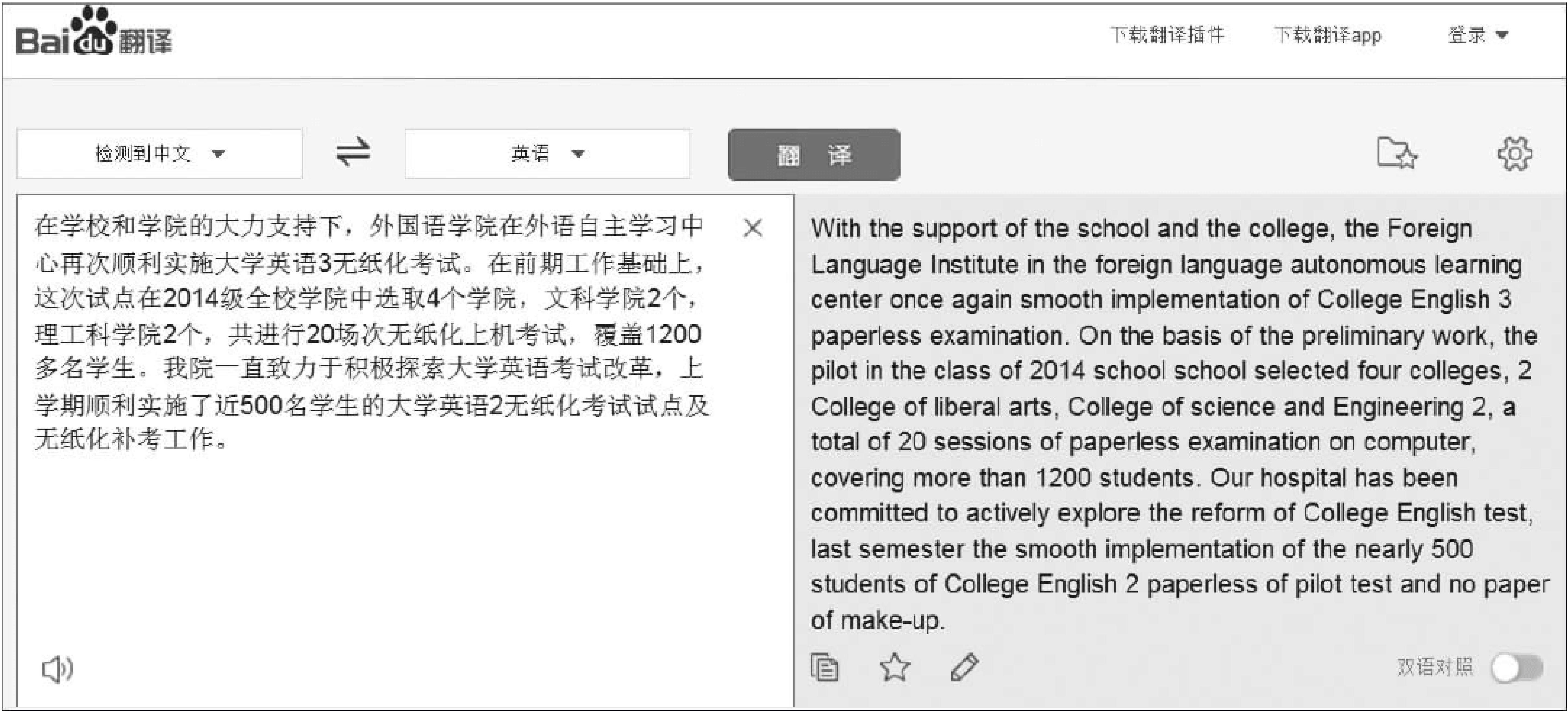


图 11-29　百度翻译实例

5）利用百度寻找范文

写应用文的时候，找几篇范文对照着写，可以提高相应的工作效率。

（1）找市场调查报告范文。市场调查报告的网页有几个特点：第一是网页标题中通常会有"××××调查报告"的字样；第二是在正文中通常会有几个特征词，如"市场"、"需求"、"消费"等。于是，利用 intitle 语法，就可以快速找到类似范文。例如检索式为：市场 消费 需求 intitle:调查报告。

（2）找申请书范文。申请书形式多样，常见的比如入党申请书。申请书有一定的格式，因此只要找到相应的特征词，问题也就迎刃而解。比如入党申请书最明显的特征词就是"我志愿加入中国共产党"。例如检索词为：我志愿加入中国共产党 入党申请书。见图 11-30。

6）利用百度寻找谜底

（1）猜谜语。有时候会遇上各种高难度的谜语，但有了搜索引擎，通常都可以在网上找到答案。搜索时只需把谜面和"谜底"作为关键词搜索就可以了。例如检索内容为：眼皮上落着一只苍蝇 谜底。

（2）解难题。除了猜谜语外，还会遇到一些类似福尔摩斯探案之类的智力题。有这么一个推理题："一个人在朋友家吃饭，问朋友这餐吃的是什么肉？朋友说是企鹅肉，他



图 11-30　百度范文搜索实例

就号啕大哭自杀了”。为什么呢? 搜一下。这个题目中的特征词串是“企鹅肉”和“自杀”,再加上问题答案的特征词“答案”,就可以快速找到结果了。再比如,微软招聘曾有一个著名的题目:下水道的盖子为什么是圆的。也可以用搜索引擎找其他人五花八门的解答。例如检索式为:企鹅肉 自杀 答案。

7)利用百度寻找医疗健康信息

互联网上有大量的健康和疾病治疗方面的资料信息,“他”就像一个超级大夫,才高八斗,学富五车,关键是要看用户怎么去向“他”咨询。

(1)根据已知疾病查找治疗方式。这类资料通常有这样的特点,在标题中会注明疾病的名称,同时会有诸如“预防”、“治疗”、“消除”等特征性关键词。于是用疾病名称和特征性关键词,就可以搜到相关的医疗信息。例如,消除青春痘、预防口腔溃疡。

(2)找专业疾病网站。对于某些大型的综合类疾病,如心脏病、癌症、艾滋病等,也可以先用搜索引擎查找这类疾病的权威专业网站,然后到这些专业网站上求医问药,获取有关知识。就是用疾病名称作为关键词搜索,搜索引擎通常会把比较权威、质量比较高的网站列在前面。例如,艾滋病。见图 11-31。

(3)根据症状找疾病隐患。经常还会有这样的需求,已知身体不舒服的症状,希望知道可能的疾病隐患是什么。这也可以通过搜索引擎解决问题,一般的疾病介绍资料,通常会有疾病名称、疾病症状、治疗方法等部分。我们描述的症状,如果和某个网页中的疾病症状刚好符合,搜到这样的网页,使用疾病名称也就知道了。做这类搜索的关键是,如何把症状现象用常用的表达方式提炼出来。例如,经常打嗝。



图 11-31　专业疾病搜索实例

8）利用百度寻找网上购物信息

直接搜索产品即可购买，直接搜索商品相关信息即可获得对应产品相关购物网站信息。例如，在搜索框输入：金庸作品集，即可获取相对应的购物网站信息。单击相应链接即可直接购买。见图 11-32。



图 11-32　百度网购搜索实例

11.2　搜狗搜索引擎的信息检索与利用

自 2003 年以来，搜狗先后推出搜狗搜索、搜狗输入法及搜狗浏览器等战略级产品，并开创了“输入法、浏览器、搜索”三级模式，成为行业追赶者的唯一成功模式。2010 年搜狗

从搜狐分拆运营,从一个部门成长为一个公司;2013 年搜狗引入腾讯的战略投资,合并了腾讯搜搜等业务。搜狗是中国互联网领先的搜索、输入法、浏览器和其他互联网产品及服务提供商。从 2004 年 8 月搜狐公司推出全球首个第三代互动式中文搜索引擎——搜狗搜索以来,历经十载,搜狗搜索已发展成为 PC 端搜索三强(Google、Baidu 与 Sogou)之一,移动搜索排名第二。根据艾瑞咨询 2015 年 8 月数据,搜狗 PC 用户规模达 5.21 亿人,仅次于腾讯,成为中国第二大互联网公司。搜狗搜索结合腾讯独家资源,打造微信搜索,上线本地生活、扫码比价、微信头条等独有服务,第一次实现了真正的差异化竞争,一方面不断拉大与跟随者的距离,另一方面不断冲击榜首位置。

1. 搜狗搜索入门

1) 开始第一次搜索

在搜索框内输入要查询的内容关键词,敲击回车键(或者单击搜索框右侧的搜狗搜索按钮)后就可以获得想要的搜索内容,无须下载、安装融合插件。例如,想查找好看的电影,在搜索框内直接输入好看的电影,敲击回车键或者单击“搜狗搜索”按钮,就可立即获得优质的结果。见图 11-33。

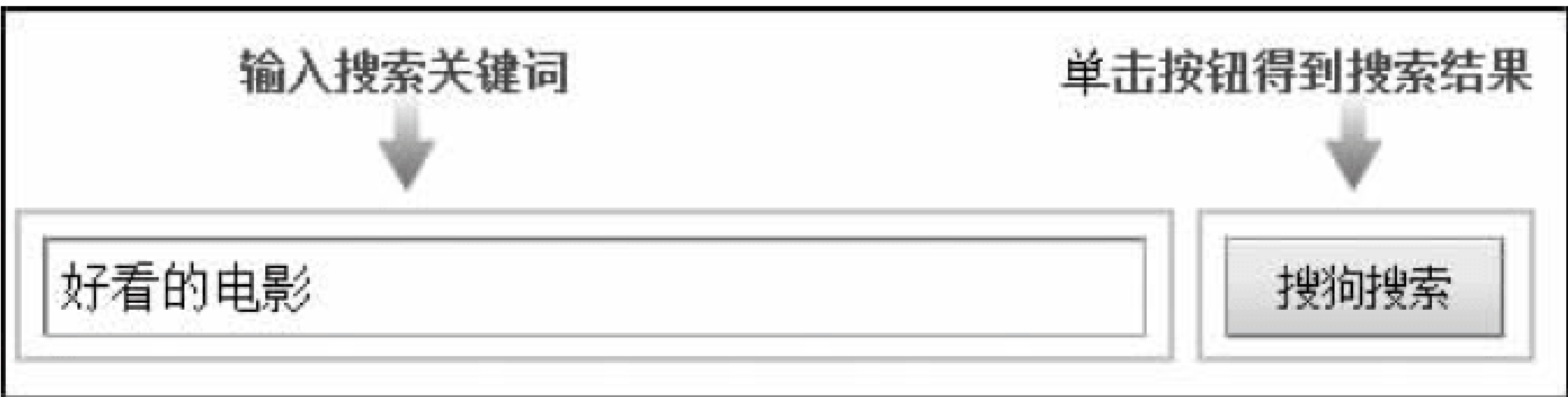


图 11-33　使用多个词语并用空格分开的检索实例

如果您想得到更精确的搜索结果,只需输入更多的关键词,并在关键词之间用空格分开。例如,搜索“中国 北京 天安门”,这样会比直接搜“中国北京天安门”结果要好。见图 11-34。

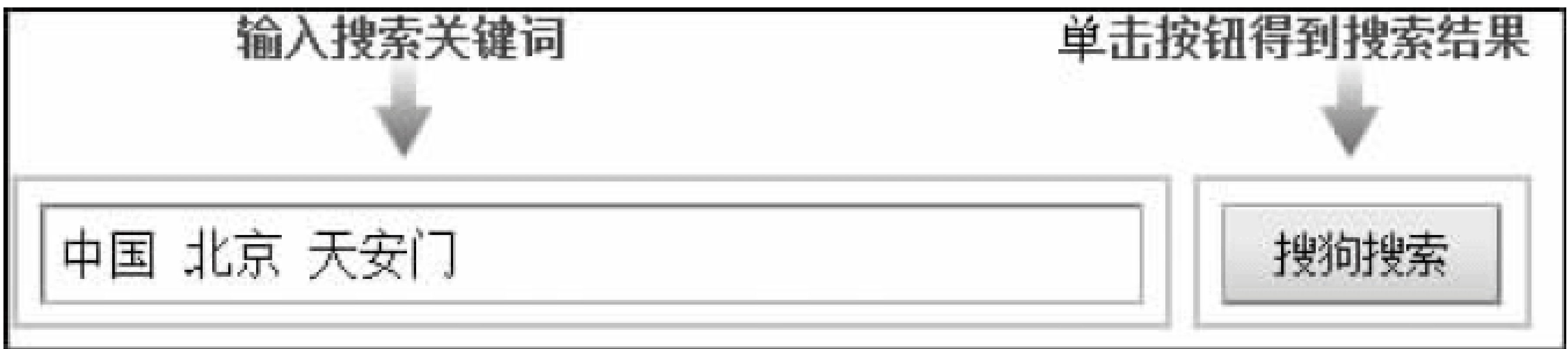


图 11-34　使用单个检索词语搜索实例

2) 用户搜索结果界面的含义

对于如何浏览搜索结果页,用户可能忽略了 50%的搜索结果界面信息。每个带下划

线的蓝色行都是用户搜索词找到的搜索结果。搜狗很贴心地把最相关的匹配项放在最前面,单击就可以打开对应的网页。以下的示例图 11-35 可以帮助用户了解搜索结果页中所有的结果元素和工具,分别用七个部分进行说明。

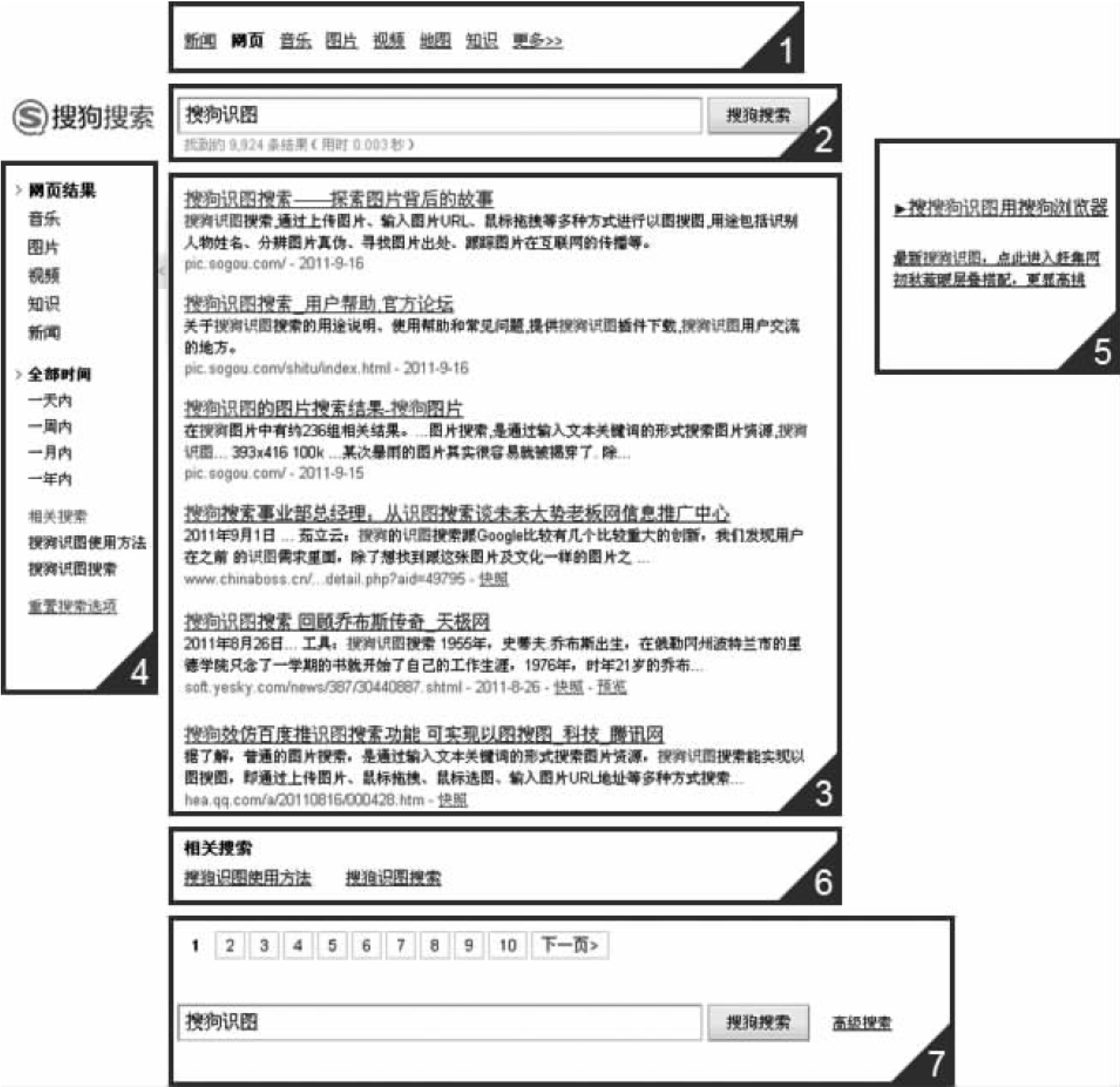


图 11-35 搜索结果实例图

第一部分:信息的分类。图 11-36 是结果信息页的第一部分即信息分类标题部分,信息分类标题是对信息类别进行的总体分类,也就是分类搜索的意思。分类标题有新闻、网页、音乐、图片、视频、地图、知识等。选择这些类别可以更加精确地搜索用户需要的信息范围,单击“更多”可获取更多的搜狗产品。

第二部分:搜索框。搜索框是搜索引擎接收用户搜索词的接口,用户输入检索词后

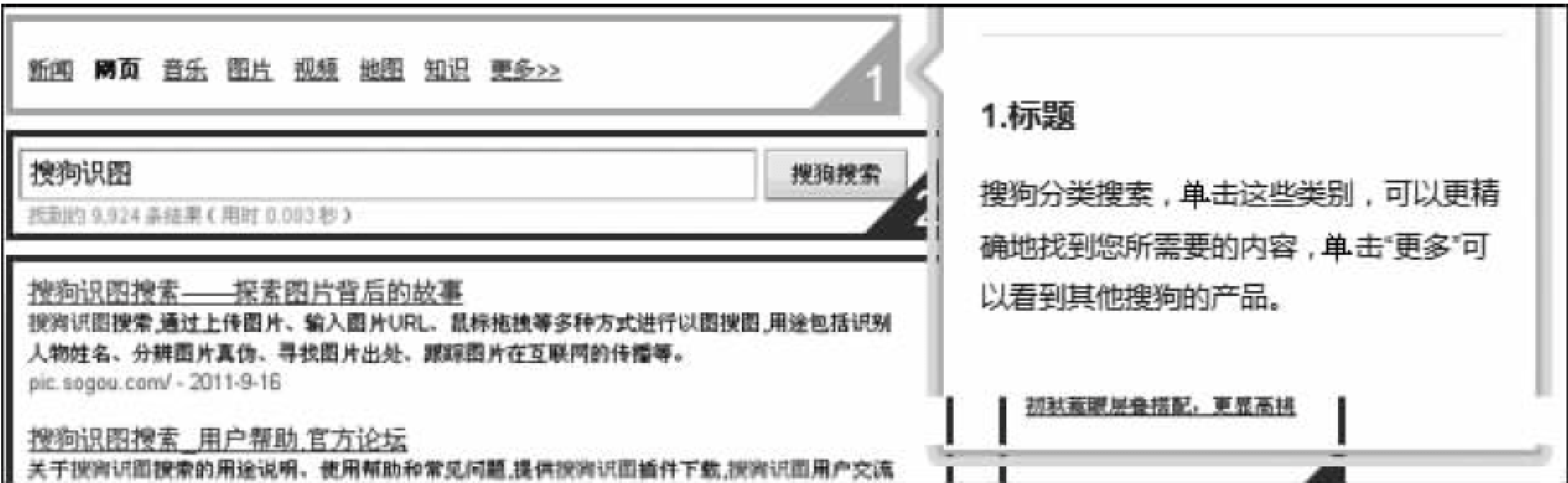


图 11-36　结果页第一部分“信息分类标题”实例

按回车 Enter 键或单击“搜狗搜索”按钮即可。其中在用户输入搜索项的词语时，在搜索框位置系统能动态提示与用户搜索词相关的最热门搜索，以提示用户评估或修正自己的检索词，以便于获得最佳的搜索结果。见图 11-37。

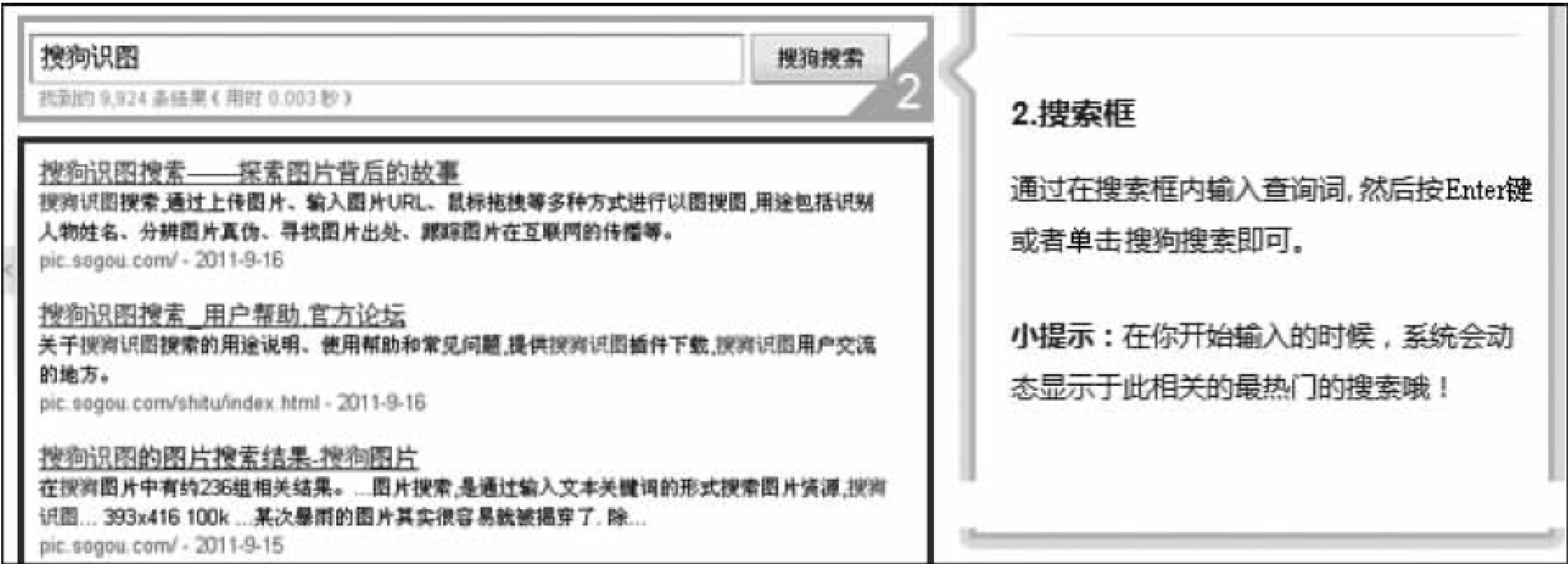


图 11-37　结果页第二部分“搜索框”实例

第三部分：搜索结果。在搜索结果内容中，依据结果内容与用户检索词的相关性程度对反馈的信息进行排序。搜索结果项包括查询反馈信息的标题、摘要、网址、快照、网页的网址及其更新时间等。见图 11-38。

第四部分：选择工具与条件过滤。选择工具内容包括网页结果的音乐、图片、视频、知识与新闻，便于用户对搜索结果的信息类别进行限制，也就是限定为特定信息类别的查询，默认信息类型为“网页”。时间筛选：可以选择搜索最新或某一段时间内的信息。相关搜索：如果首次搜索，有可能拟定的检索词不是很精确而达不到理想的反馈结果，这时可以参考其他网友的搜索方法，即相关搜索以提高结果质量。重置搜索结果：清楚用户之前的筛选条件，开始新的查询。见图 11-39。

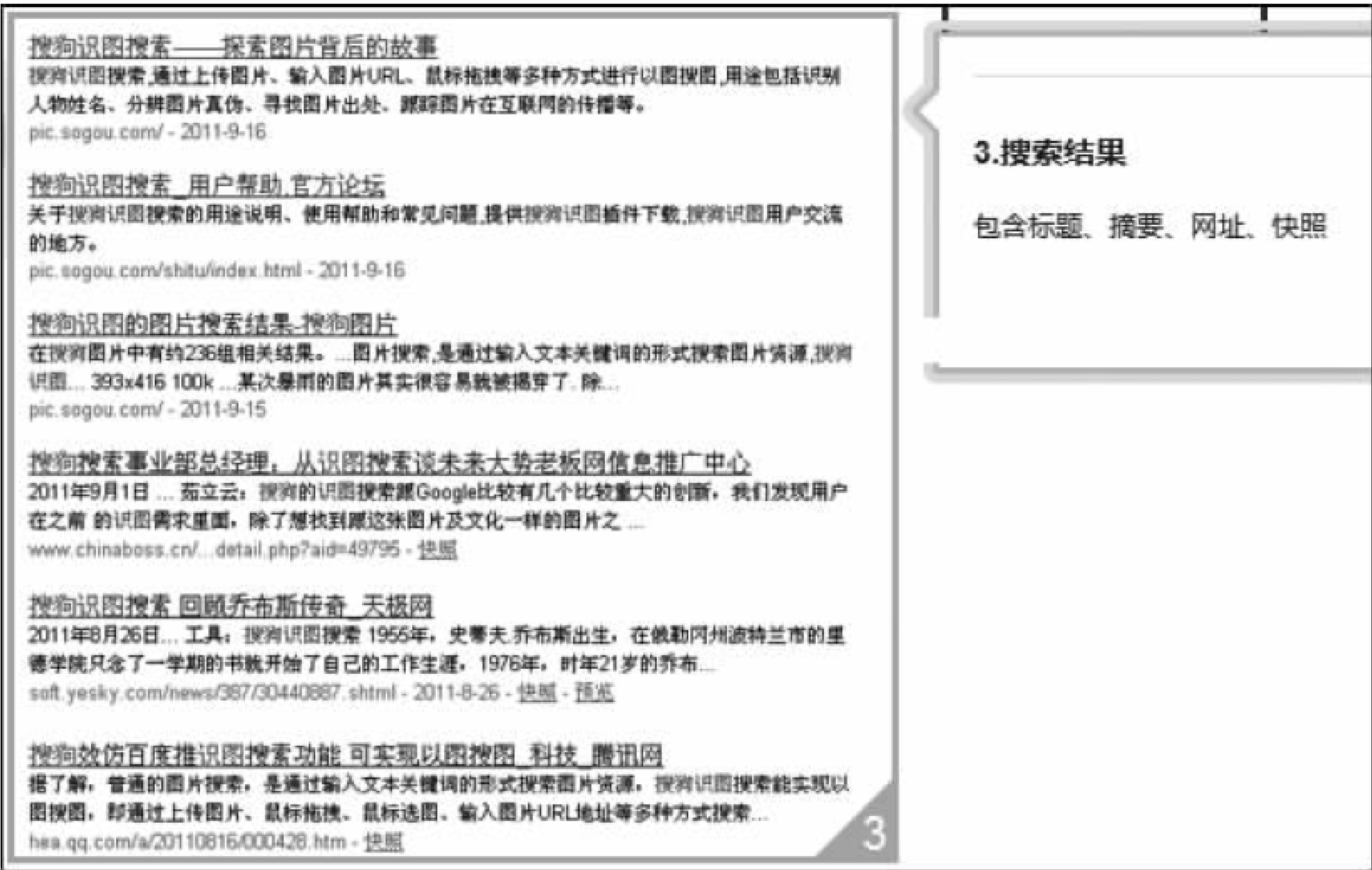


图 11-38　结果页第三部分“搜索结果”实例

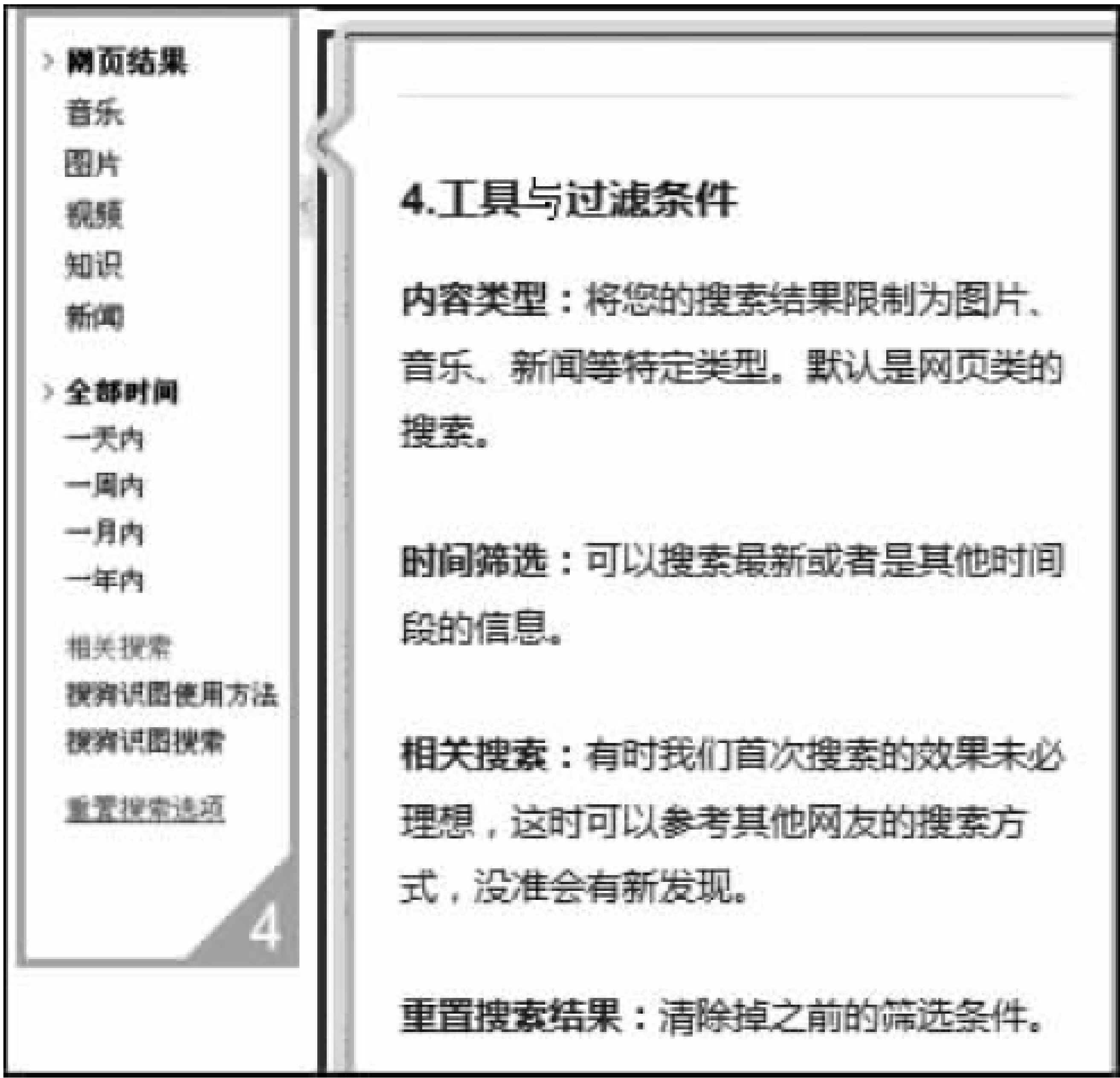


图11-39　结果页第四部分“工具与过滤条件”实例

第五部分：广告。这些广告与用户的搜索内容相关，为用户需要查询的内容提供有价值的参考。如果用户希望展示自己的网站，也可以进一步了解相关广告内容与事项。见图 11-40。

第六部分：相关搜索。参考其他网友的相关搜索可能会获得更好的搜索结果。见图 11-41。

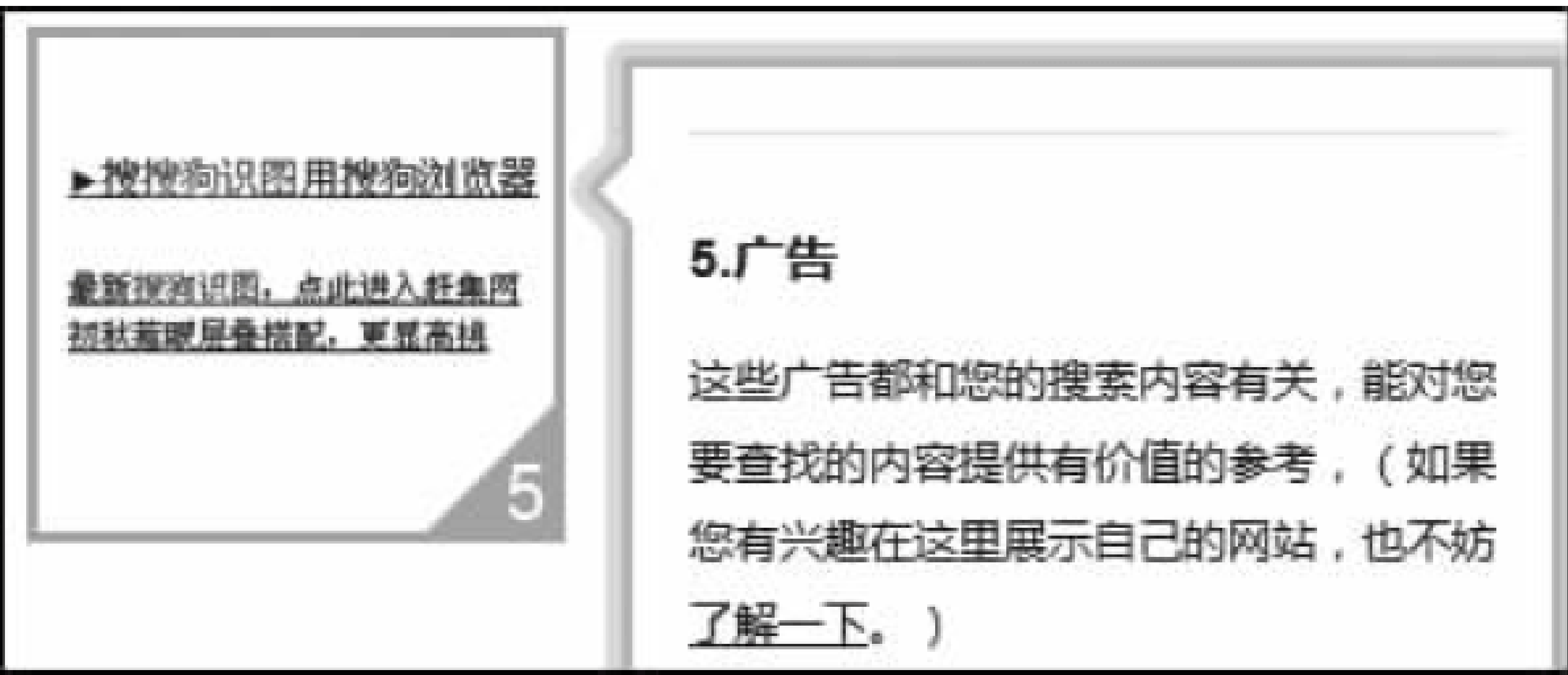


图 11-40　结果页第五部分"广告"实例



图 11-41　结果页第六部分"相关搜索"实例

第七部分：网页底部。网页底部有更多的结果显示（用页码序号提示）和翻页导航。见图 11-42。



图 11-42　结果页第七部分"网页底部"实例

3）删除搜索历史

很多用户在搜索时总是会自动填充以前搜索过的内容，不知如何解决，其实这是网页浏览器的一项基本功能。需要进入 IE 浏览器的相关菜单进行设置：如果您使用 IE4.0 浏览器，则可通过"查看→Internet 选项→内容→自动完成→清除表单→完成"进行设置；如果您使用 IE5.0 及以上版本的浏览器，则由"工具→Internet 选项→内容→自动完成→清除表单→完成"进行搜索历史删除。见图 11-43。

如果希望 IE 浏览器以后不再记录查询过的内容，请在"自动完成"设置页面内把"表



图 11-43　IE 浏览器中“删除搜索框历史”的操作实例

单”前的选项钩去掉。如果您使用的是搜狗浏览器，在工具栏的“清除浏览记录”中便可以轻松删除，见图 11-44。



图 11-44　搜狗浏览器“删除搜索框历史”操作实例

4）不能正常访问搜狗引擎的常见解决办法

(1) 确定是否其他网页也无法访问，以排除网络原因。

(2) 重启一次浏览器，并尝试重新连接到 Sogou。

(3) 重启计算机，清除浏览器缓存并删除 Cookie。

(4) 使用防火墙、代理服务商或防病毒程序。

(5) 清除计算机的 DNS 缓存清除 Hosts 文件。

2. 搜狗搜索技巧

1）如何选择查询词

最基本、有效的查询技巧，就是选择合适的查询词。以搜索引擎容易分辨的词语来查询，能够大大提高查询效率。

(1) 简单明确：每个查询词都应该使目标更加明确，尽量减少无关重复的词语。例如，

✖“简简单单不复杂又好听的网名”的查询词太长，完全符合条件的结果可能较少。

✔“简单的网名”效果更好。

检查您有没有把自己的想法以对话的方式输入查询词。例如，

✖搜索“我想看暑假最多人喜欢的电影”，搜索引擎不会理解，查询词太长。

✔搜索“暑期 热门 电影”效果更好。

(2) 使用网页中会出现的语言。尽量使用网页上可能出现的词。例如，

✖“很多人喜欢的来电声音”。

✔“来电铃声”或“手机铃声”。

以上比较好的查询词采用的都是网络中比较常用的词汇，更有利于得到优质结果。多留意网页上会出现的词，并且去猜测信息的表达方式并提取关键词，会大大提高搜索的准确率。

2) 高级搜索的常用语法

(1) 精确匹配（“”）。利用双引号可以查询完全符合关键词字串的网站。例如直接输入热门游戏，会返回“热门网络游戏”、“热门小游戏”、“游戏下载”等内容，如果输入“热门游戏”(用双引号进行了精确匹配)，搜狗就会严格按照该检索词的完整形式查找内容，不做任何拆分。

(2) 在特定网站内搜索(site：)。见图 11-45。

site语句查看收录情况

除了基本站内查询外，搜狗还为站长和网站管理员们提供了更加强大的功能，使用的时候不加关键词，只需要输入“site:站点域名”，就可以查找你的网站在搜狗的收录里。提供的功能包括：

收录查询：输入“site:网站域名”，可以查到站点在sogou的收录情况，如果博客地址，就可以知道你的博客在搜狗的收录情况，比如： 17173 Good Feel MC的共享空间

域名后缀：你可以查看一个域名或者子域名下的内容，结果按照重要性排序。比如 “site:cn"是指在所有网站域名最后为cn的收录情况，您可以试着查询一下这些网站：国内公司 政府机关 教育机构 香港地区网站

支持https：所有以https://开头的网站链接也都能使用站内查询查找，比如： 招商银行一网通主站

端口查询：大部分网站使用80端口，不需要特别指定，但有些开在其他端口，这个时候使用“site:站点域名:端口号”进行查询即可。

海里数据支持：搜狗的站内查询内容丰富，针对您的查询站点，提供全面的数据

图 11-45　Site 语句查看收录情况说明

如果想知道某个站点中是否有自己需要找的东西，可以使用 site 语法，其格式为：查询词＋空格＋site：网址。例如只想看搜狐网站上的财经新闻，就可以这样查询：财经 site：sohu.com。搜狗还支持多站点查询，多个站点用“|”隔开，“site：”和站点名之间，不要带空格。例如检索式为：site：www.sina.com.cn |www.sohu.com。

(3) 在特定的网页标题中搜索(intitle:)。如果需要把搜索范围局限在特定的网页标题中,可使用 intitle 语法,其格式为:查询词+空格+intitle:网页标题所含关键词。例如,找周杰伦的新歌,则检索表达式为:新歌 intitle:周杰伦。

(4) 特定文件搜索(filetype:)。如果不是想搜网页内容,而是想找某一类的文件,filetype 语法可以解决这个问题。其搜索语法为:查询词+空格+filetype:格式,格式可以是 DOC、PDF、PPT、XLS、RTF、ALL(全部文档)。例如检索式为:市场分析 filetype:doc,其中的冒号是中英文符号皆可,并且不区分大小写。filetype:doc 可以在前也可以在后,但注意关键词和 filetype 之间一定要有空格。例如,filetype:doc 市场分析。filetype 语法也可以与 site 语法混用,以实现在指定网站内的文档搜索。例如,site:www.cau.edu.cn|www.tsinghua.edu.cn filetype:all 中国,表示的含义是在中国农业大学和清华大学网站内搜索有关"中国"的文档。

3) 高级搜索功能

如果对搜狗的各种查询语法不熟悉,可以使用集成的高级搜索功能,方便实现高级搜索语法功能。高级搜索的各项功能如图 11-46 所示。

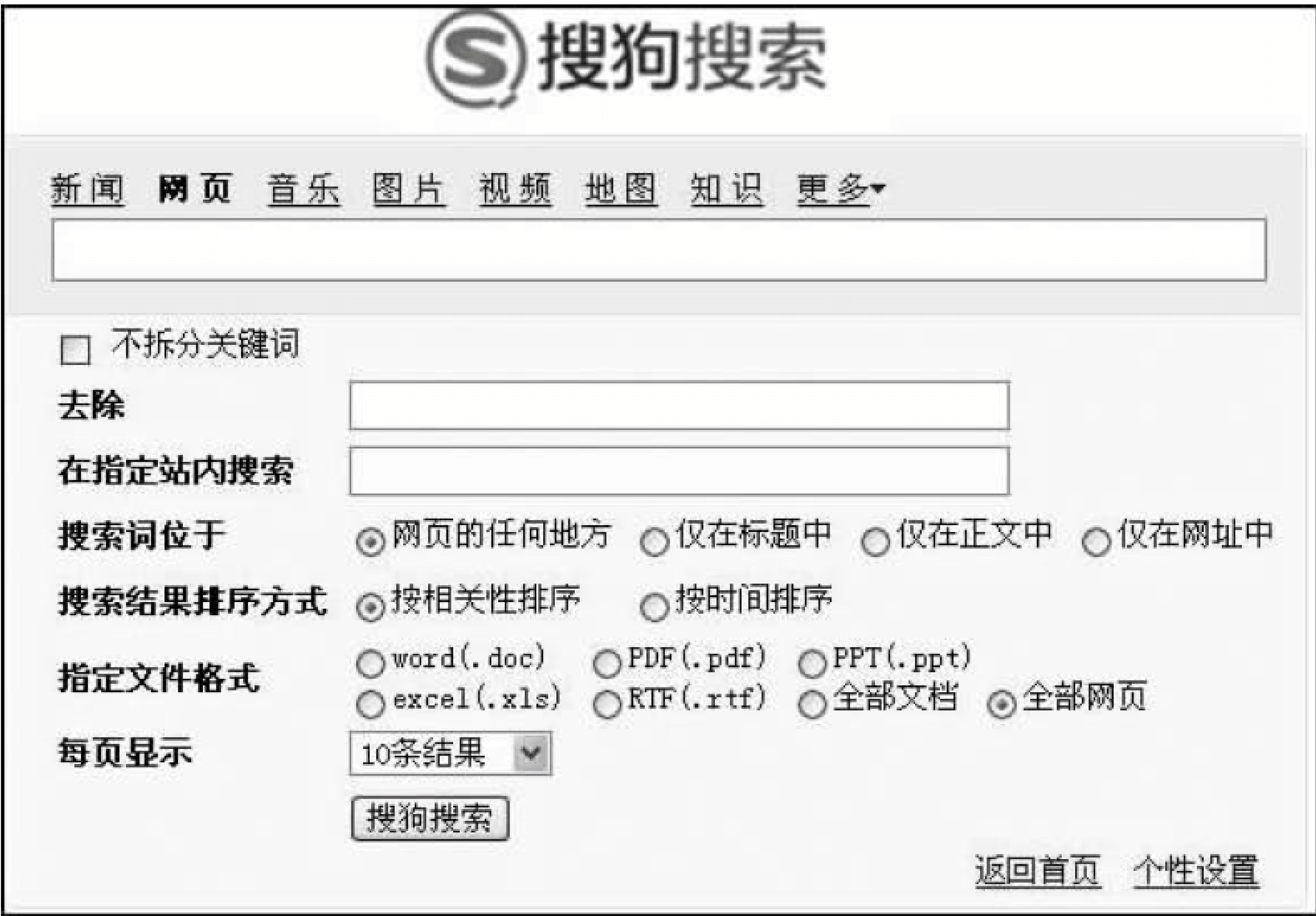


图 11-46 搜狗高级功能

(1) 去除:如果想要避免搜索中包含某些内容,可以将需要避免的内容填在框中。例如,需要查询"仙剑奇侠传",希望查看其游戏方面的信息,但搜索结果中包含较多该查询词的电视剧内容,则只需要在搜索框中输入"仙剑奇侠传",在"去除"框中输入"电视剧"。

(2) 在指定站内搜索：比如只想看搜狐网站上的新闻，就可以在顶端搜索框中输入"新闻"，在"指定站内搜索"框中输入 www.sohu.com。

(3) 搜索词位于：可以把搜索范围局限在特定的网页标题、网页正文、网页网址当中，使用时只要选中需要的范围即可。

(4) 搜索结果排序方式：按相关性排序可以让与搜索词匹配程度最高的结果排在前列，按时间排序则是按搜索结果的时间顺序由新至旧排列。

(5) 指定文件格式：如果要查询某一类格式的文档，直接在这一栏勾选想找的文档类型即可。

(6) 每页显示：修改每一页结果的显示数量，搜狗支持每页显示 10 条、20 条、30 条等结果显示。

4) 个性设置

用户可以根据自己的搜索习惯，在个性设置界面中改变搜狗默认的搜索结果显示条数和搜索结果打开方式。搜索结果显示条数设置：当用户想一次性浏览大量信息时，可以在此修改每一页结果的显示数量，搜狗支持每页显示 10 条、20 条、30 条、50 条或 100 条结果，默认的是每页 10 条结果；搜索结果打开方式：可以设置单击搜索结果是否在新窗口打开，默认的是打开新窗口。见图 11-47。



图 11-47　搜索引擎的个性设置界面

3. 搜索框提示

当开始向搜索框输入拼音或者文字时，搜狗马上开始推测用户想要输入的内容，并提供实时建议。例如，用户输入"xiaosh"或者"小说"就会出现如图 11-48 所示的提示。如果手气不错，用户不需要输入全部的检索信息，就可以通过使用箭头键或鼠标选择所需要的提示信息。而且搜狗的提示信息都是根据信息的热门程度来预测的，用户也可以看看最近的相关信息热搜榜。

1) 拼音提示

如果觉得切换中文输入法太麻烦，或者只知道某个词的读音而不知道字形，用户只要



图 11-48 搜索框提示的实例图

输入查询词的汉语拼音，搜狗就能在搜索框中给出最符合的汉字提示供用户选择。用户也可以直接按 Enter 键，拼音提示自动会出现在搜索结果上方。例如输入 qinghua 后提示为“您是不是要找：清华”。见图 11-49。



图 11-49 拼音提示的实例

2）错别字提示

我们在打字输入检索词时经常会输入一些错别字，导致搜索的结果根本是不需要的信息。有了搜狗错别字提示功能，这个问题就迎刃而解了，被打错的字会显示在结果上方，并且直接显示正确字形的搜索结果。例如输入青华大学，会提示“您是不是要找：清华大学”。见图 11-50。

4. 搜狗信息搜索服务产品

搜狗提供一系列的搜索服务产品，主要有以下几种。



图 11-50　错别字提示的实例图

（1）新产品推荐。例如搜狗明医、搜狗知乎搜索、搜狗软件下载、搜狗微信搜索等。见图 11-51。



图 11-51　搜狗新产品推荐

（2）搜狗产品。产品丰富，包括网页、音乐、视频、图片、学术、文档、论坛等 26 种。见图 11-52。



图 11-52　搜狗搜索产品推荐

（3）搜狗桌面。项目产品包括搜狗高速浏览器、搜狗拼音输入法和搜狗壁纸。见图 11-53。

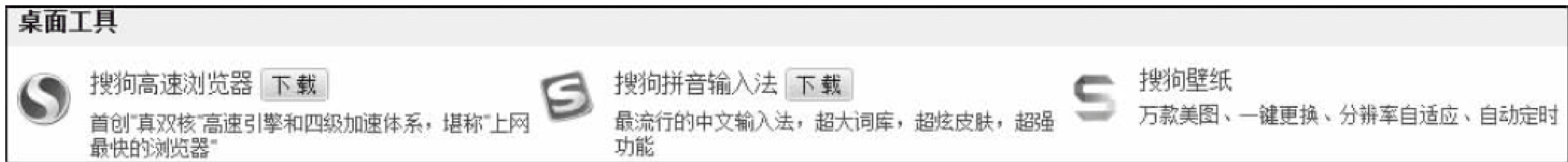


图 11-53　搜狗桌面工具

（4）手机软件。手机端产品有搜狗手机助手、手机输入法、搜狗语音助手等 10 类。见图 11-54。

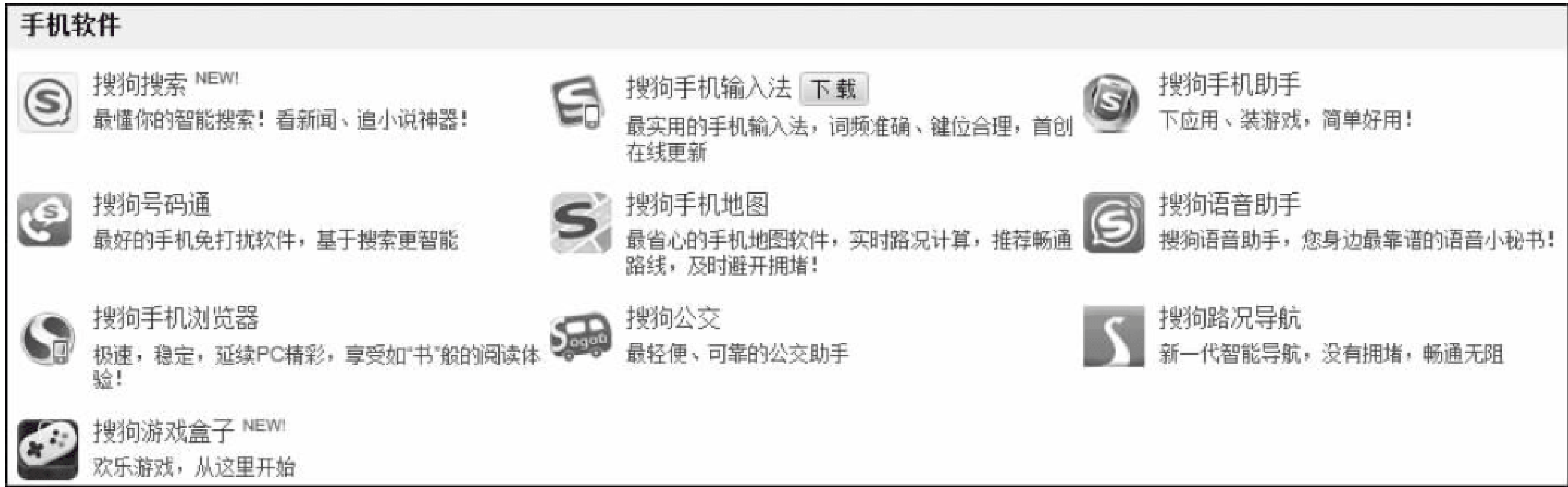


图 11-54　搜狗手机软件服务产品

11.3　Google 搜索引擎的检索应用

1. Google 概述

Google 网络搜索引擎是目前世界上发展最快、规模最大、网络用户量最多的大型搜索引擎。Google 创建于 1998 年 9 月，创始人为 Larry Page 和 Sergey Brin。Google 的使命是整合全球信息，使人人皆可访问并从中受益。Google 允许以多种语言进行搜索，在操作界面中提供多达 132 种查询语言。Google 搜索引擎的主要搜索服务有网页、图片、音乐、视频、地图、新闻、问答等搜索服务产品。Google 中文版搜索主界面如图 11-55 所示。

“谷歌”是 Google 公司针对海外中文用户市场而起的唯一一个中文名字，谷歌在发音上与 Google 相似，同时也融合了中国传统文化的含义。谷歌的意思就是以谷为歌，是播种与期待之歌，亦是收获与欢愉之歌。在搜索信息时如果选择英文搜索，单击右下角的



图 11-55　谷歌引擎中文搜索用户界面

English 链接即转到英文界面，英文搜索界面和中文界面基本一致，英文搜索用户界面如图 11-56 所示。



图 11-56　谷歌英文搜索

Google 搜索引擎以其使用简单、干净简洁的用户检索界面，检索结果与用户查询需求的相关度高，提供的搜索关联业务服务产品丰富等优势，赢得了越来越多因特网用户的广泛认同。谷歌搜索引擎每天需要处理两亿多次网络用户的搜索请求，数据库存有 30 亿个 Web 文件，提供常规初级搜索和高级搜索两种功能。

2. 便捷实用的 Google 翻译功能

对于大学生的探究性和研究性学习而言，查询与获取前沿性、质量高的外文资料可以帮助开拓思路和及时了解国际领先成果，避免人力和时间上的浪费。Google 的多语种翻译功能为外文资料检索带来了极大的方便，它应用计算机智能翻译技术，打破了语言上的障碍，甚至可以查到词典上没有的生词，Google 的翻译页面如图 11-57 所示。

如果学生在搜索某一外文主题资料时不知道相应的英文表述，或在阅读外文资料时



图 11-57 谷歌翻译主界面

遇到生词影响进度,可用 Google 的"中英文词典"来解决这些难题。只要在 Google 的搜索框中同时输入"翻译"或"fy"与要翻译的中文或英文词汇,在返回的结果网页的最上方就可以找到相关的翻译信息。如果用户使用 Google 搜索外文网站,会惊喜地发现搜索结果页面多数网站链接后都出现了"[翻译此页 BETA]"链接,单击它就可以看到 Google 自动翻译的中文页面,该网页翻译服务实现了中文到英文的智能翻译。

3. 快捷有效的 Google 特殊操作符搜索

在进行信息检索的过程中,很多用户都遇到过这样的问题:检索结果中有很多与检索词无关或没有学术或研究性价值的链接,而有用信息则被淹没其中。在输入检索条件时使用一些特殊操作符,可以起到事半功倍的效果。

(1) 用"filetype"搜索指定信息文档类型。Google 支持 13 种非 HTML 文件的搜索,包括 Microsoft Office 系列文档(doc、ppt、xls、rtf),Adobe 公司的 pdf 文档和 swf 文档等,还支持 jpg 图片格式的文档。使用"filetype"来搜索指定类型的文档,可以大大拓宽 Google 用户在网上获得信息的目的性。如果用户想查找有关虚拟现实技术方面的学习课件,只需搜索"filetype: ppt 虚拟现实技术",搜索结果中出现的链接将都是 ppt 文档。Google 可以为用户提供不同类型文件的"HTML 版",方便用户在未安装相应应用程序的情况下阅读各种文件内容,用"HTML 版"阅读能帮助用户防范某些类型文档可能带来的病毒。

(2) site 操作符。用"site"限制在某个网站或网站的某个网页内进行搜索。互联网上有许多网站本身并不具备网站搜索功能,想要在这些网站中查找一些资料十分费力,这时可以利用 site 操作符对这个网站进行内部搜索,简便地找到所需的资料。例如,某个搜索

用户想了解桂林电子科技大学 2016 年的研究生招生信息，只需在搜索框中输入“site：www . guet. edu. cn 研究生招生”就能快捷地找到所需网页内容。

（3）In-系列搜索指令

In-系列搜索指令是 Google 搜索中最重要的“位置关键词”查找方式，通过 intitle、inurl、intext 三个搜索指令来指定关键词的位置，可以分别查找在标题、链接、正文包含搜索关键词的网页结果。对于目标明确的搜索者来说，In-系列搜索指令往往最为简洁，能够有效简化搜索结果，提高搜索精确度。

① Inurl 链接搜索。Inurl 操作符可以限制所搜索关键词包含在 URL 链接中。任何网站的 url 都不是随意设置的，url 链接通常和网页的内容有着密切的相关，利用这种相关性可以缩小搜索范围，快速找到所需信息。比如，提供书籍下载的 url 一般包括 book、ebook、shu、shuji 等，而与软件相关的会使用 soft、software、ruanjian 等。平时注意观察网页的 url，就能总结出不同资源的常用 url。如果要查找数据挖掘方面的资料，可以使用“inurl：book 数据挖掘”这个检索表达式，就可以搜索到很多相关书籍的网站。

② Intitle 标题搜索。intitle 操作符可将搜索的关键词包含在网页的标题中，网页在设计时一般都会把网页的关键内容用简明的语言显示在网页的标题中。利用 intitle 操作符对网页的标题栏进行搜索，一般都会找到相关率比较高的专题性页面。例如，搜索中国知网的相关信息，只需输入“intitle：cnki”即可查询到所需网页。

③ Intext 正文检索。与标题搜索相比，正文检索的搜索目标更明确，而且适合于一次性搜索同一主题的不同分支内容。例如，如果想要找到高血脂的病因及其治疗方面的信息，就可以利用：“intext：高血脂＋病因＋治疗”来得到理想的搜索结果。

4. Google 信息检索实用功能

（1）目录检索。Google 的分类网站目录划分明确，信息集中，大学生应养成首先考虑在相关主题网站上查找所需信息的习惯。查找专题网站，可以按学科主题进行浏览，Google 使用的分类目录采用了 ODP（公共网页目录）规范。打开网页目录，进行分类浏览，可以查看依照性质和内容分类的由世界各地义务编辑人员审核挑选的网页。在检索时选择在某一目录门类中进行搜索，往往要比同类搜索引擎有更高的命中率和检索效率。

（2）使用偏好。单击 Google 搜索按钮右侧的“使用偏好”链接，可以通过使用偏好功能轻松设置用户的个性检索。设置方法如下：如果在“界面语言”中选“中文简体”，打开的页面语言就是中文简体。如在“搜索语言”中选“中文简体”，Google 就只会在简体中文网页中进行搜索。建议选中“开启新视窗以显示查询结果”一项，这样单击搜索结果时会打开新的窗口。使用偏好设置还允许用户定制搜索结果页面所含信息条目数量，可从 10

到 100 条任选，还可以选择是否使用汉字简繁体转换，最后单击存储使用偏好，就可以将本次设置的格式套用到以后的搜索中。

（3）地图搜索与地图导航。与其他搜索引擎相比，Google 有功能最强大的地图搜索功能（包括二维地图、立体地图、全景地图等）。单击谷歌地图可以自动跳转到所在地地图，并可以在相应的搜索栏输入要检索的地图以及乘车信息等内容，并同步进行地图位置导航。Google 地图搜索实例如图 11-58 所示。

（4）图片搜索。Google 也是互联网最好用的图像搜索工具，单击 Google 首页的图像检索模块，在关键词输入栏内输入关键词，就可以找到需要搜索的图片缩图，而且可以查看原始图片及查找出该图片的出处。除了 Google 提供的专门图片搜索功能外，还可以组合使用一些搜索语法，以达到准确图片搜索的目的。其中一种是利用专门提供图片集合的网站，通常会把图片放在某个专门的目录下，如/gallery、album、photo、image 等，这样就可以使用 inurl 语法迅速找到这类目录。另一种是提供图片集合的网页，通常在标题栏内会注明某个图片集合，可以用 intitle 语法找到这类图片，还可以用 site 语法指定所提供图片的站点。图片搜索界面如图 11-59 所示。

（5）音像资料搜索。搜索 MP3 可以用 inurl 语法搜索，也可以用网页标题 intitle 语法搜索音像资料。例如搜索"时间都去哪了"这首歌，则搜索式为：inurl：mp3 时间都去哪了。例如搜索百家讲坛的电视视频节目，则搜索式为：intitle：电视节目 百家讲坛。

（6）软件搜索。在软件搜索时，直接输入软件名称下载，但这样随意下载是不安全的，供下载的软件有可能带有病毒或捆绑木马。需要用 site 语法对下载网站进行限定。搜索下载软件的 serial、number、sn 等序列号信息，直接输入关键词即可。例如搜索 winzip10.0 的注册码，则搜索式为：winzip8.0 sn。

（7）近似词搜索。如果需要搜索同义词或者近义词，需要在检索词前加"～"。例如"～elderly"可以获得包括"senior"、"older"、"aged"等内容的网页。

5. Google 高级搜索

在 Google 中，除了普通的搜索外，还可以进行高级搜索。在高级搜索界面，可以输入需要的多个检索词进行高级搜索逻辑限定，以提高信息搜索的准确性。Google 高级搜索主界面如图 11-60 所示。

对于大学生而言，要提高网络信息搜索的查询质量与查询效率，避免在检索结果中出现过多不相关信息而导致的信息噪音干扰，无论是在使用搜索引擎还是一般检索数据库时，都需要逐步形成信息"高级搜索"的基本素养。Google 高级搜索的主要功能如下。

（1）"以下所有字词"：例如直接输入分布式网络数据库系统，在检索时包括了分布

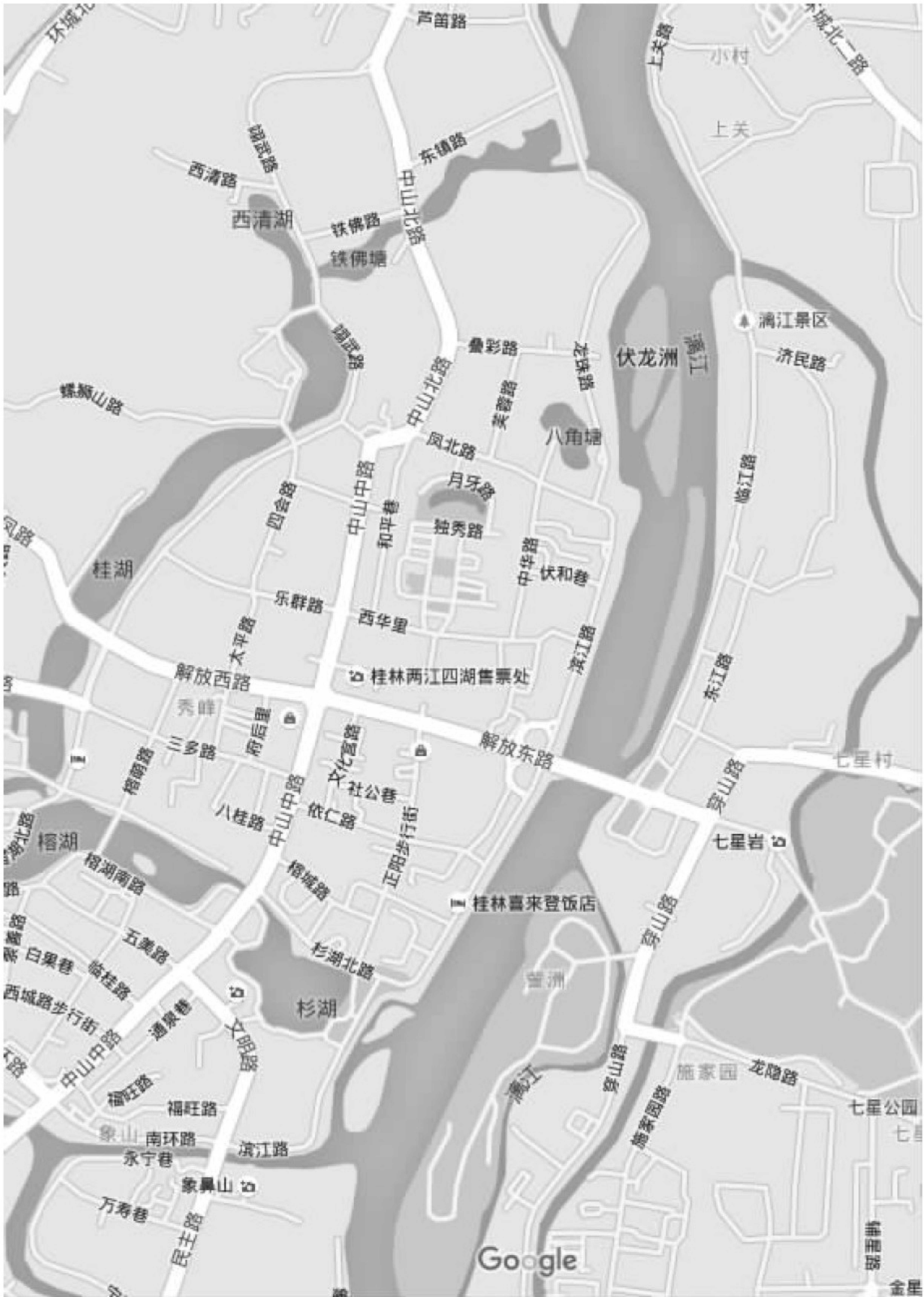


图 11-58　谷歌地图搜索实例(桂林市中心区)

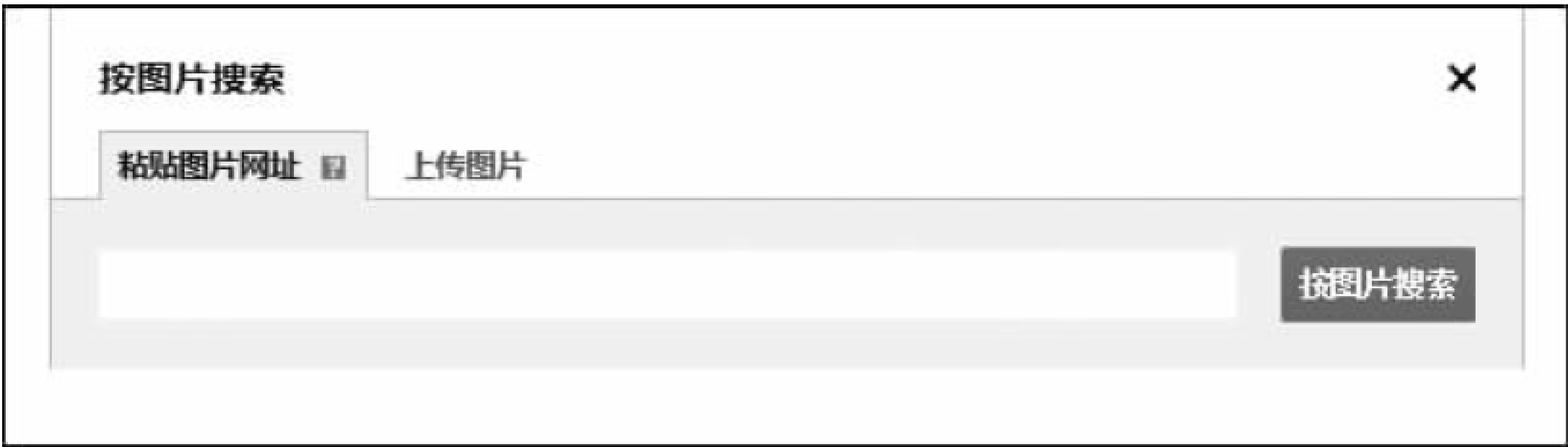


图 11-59　谷歌图片搜索主界面

图 11-60　Google 高级搜索主界面

式、网络、数据库、系统、分布式网络、网络数据库、数据库系统、分布式网络数据库系统等所有字词在内。“以下所有字词”的检索广度较高，拓展了信息检索的范围，因此信息查全率较高。

（2）“与以下字词完全匹配”：需要用双引号将检索词括起来，例如，“移动互联网”表示完全匹配检索词“移动互联网”。“与以下字词完全匹配”的检索查准率高。

(3)“以下任意字词”：在检索时需要 OR 连接，例如，批发价 OR 团购价 OR 特价，检索结果包括了商品的批发价、团购价、特价等内容。“以下任意字词”检索的查全率较高。

(4)“不含以下任意字词”：检索时在检索词前加减号(即－)，例如，－山大、－鸭梨，表示获取的信息中剔除了“山大”和“鸭梨”方面的信息。

(5)“数字范围”：在检索时在两个数字检索词之间用两个点号分开，并在数字旁添加度量单位。例如，200..300 公斤、2013..2016 年。

(6)“语言”：用户指定搜索结果网页的语言类型即查找用户所熟悉的网页语言(例如韩语网页)。

(7)“地区”：用户指定在特定地区发布的网页。

(8)“最后更新时间”：查找用户指定时间内更新的网页。

(9)“网站或域名”：搜索用户指定的网站(例如 www. guet. edu. cn)内容，或者将搜索结果限定在指定的域名范围内(例如. org 、. gov 或 . com 等)。

(10)“字词出现位置”：用户可以限定搜索的关键字词出现在整个网页、网页标题、网址或网页中链接的字词位置。

(11)“安全搜索”：用户可以设置安全搜索的等级为适中、严格或关闭，指定安全搜索用来针对色情内容的过滤等级。

(12)“文件类型”：指定所查找网页的文件格式，例如，. PDF、. PPT、. FLV、. DOC 等文件格式的网页。

(13)“使用权限”：查找不依据许可过滤，可以任意使用的网页。

(14)“个性化搜索”：包括查找类似网页或相应网页、搜索访问过的网页、在搜索框中使用通配符和自定义搜索设置。

(15)“网页快照”(cached)：帮助用户快速浏览和判定网页的大致内容，帮助查询某些链接已经不存在或者内容更换了的网页，这对于追溯一些过去的网页是有辅助作用的。

6. Google 引擎的突出特点

客观公正。Google 以其复杂而全自动的搜索方法排除了人为因素的干预，从而保证了搜索结果的客观公正性。

独特 PR 值。PR 值即网页排序(PageRank)值，是 Google 判定网页重要性的重要标准，PR 值越高说明网页的重要性程度越高，该技术也是 Google 引擎独特的专利技术。

超文本匹配分析。引擎在扫描网页文本的基础上，能够分析网页的全部内容，例如内容字体、内容分区、字词的网页位置等；同时能够分析相邻网页内容，确保搜索的返回结果有较高的相关度。

关键词接近度分析。Google 引擎不仅能够搜索出多个关键词的结果，并且能够对网页关键词的接近度进行分析，并依据接近度确定搜索结果的先后顺序，从而提高了用户评价、选择和利用信息的效率。

11.4 Infoseek 搜索引擎

相对于百度、搜狗或谷歌而言，大学生们不是很熟悉 Infoseek 搜索引擎，但是它有自己独特的搜索服务特色。比如 InfoseekChina（见图 11-61），其搜索的内容描述是英语，这对于大学生用户而言，无论是原版的英语内容学习或借鉴参考，都有很好的帮助作用。因特网在全球日益普及化的趋势使得网络信息资源也形成了全球化格局，作为网络信息检索工具也顺应了这一时代潮流。Infoseek 除了美国本土的服务版本外，也推出“InfoseekChina”、“InfoseekFrance”、“InfoseekItaly”、“InfoseekJapan”、“InfoseekUK”等多国家或地区服务版本并逐步遍及全球。



图 11-61 InfoseekChina 搜索引擎主界面

1. Infoseek 概述

Infoseek 是早期最重要的搜索引擎之一，允许站长提交网址是从 Infoseek 开始的。百度创始人李彦宏就是 Infoseek 的核心工程师之一。Infoseek 是 Infoseek 公司于 1995 年 2 月推出的万维网搜索引擎，它是一个综合网点，提供很多有用的附加服务，包括通过电子函件发送新闻、外国语检索、按地理区域的检索以及个人的金融文件夹等，Infoseek 庞大的全文数据库保证了查全率，而它独特的检索算法和一些新增加的检索功能提高了查准率，因此检索精度高，使得它由一个检索工具变成了一个强大的信息服务中心。它基于

robot 的数据发掘技术，并支持搜索结果相关性排序，并且在搜索结果中使用了网页自动摘要技术。

2. 检索方式与应用

实现分类主题一体化。在 Infoseek 的主页上既可进行分类检索，又可进行主题检索，更可贵的是 InfoSeek 的 Ultrasmart 和 Ultraseek 很好地把二者结合起来，供不同层次的用户选择使用。从人们思维的习惯角度考虑，对那些知道自己想查什么却又不能用词语确切表达出这种需求且检索经验相对较少的用户，Ultrasmart 无疑给他们提供了便利。对那些检索经验相对较丰富、对检索所花费的时间以及结果的准确度要求相对较高的用户，Ultraseek 则是很好的选择。Ultrasmart 针对网络信息自身特点的分类指南和 Ultraseek 针对全文进行索引的特性仍使网络用户受益匪浅。

1）Infoseek 目录查询

Infoseek 主页上的检索框上方有如下内容：ABC（美国广播公司）、Daytime（白天）、Late Night（夜晚）、Video（视频）、News（新闻）、Sports（体育运动）、Games（游戏）、Shop（商店）等。例如 ABC news（美国广播公司新闻）的分类栏目有 Good morning America（早安美国）、World news tonight（今晚世界新闻）、prime time live（全盛时期生活）、lightline（轻线）、World news now（实时世界新闻）等。

InfoseekChina 的分类有自己的特色，例如图 11-62 所示的主题目录有 Infoseek 中国站点（包括头条新闻、科技新闻、娱乐新闻、饮料新闻、旅游新闻等）、中国新闻媒体（包括业务、地区、娱乐、运动、全国、科技、博客与报告等）、中国站点行业（航空防卫、工业品、农业、保险、汽车、国际贸易、银行等）、交易投资（深交所、上交所、中国香港交易所、贸易与投资）等。

InfoseekWWW 页面查询的结果是，每一记录最上面一行是文件名字以及超文本文件与其他资源的接口；接着是对文件的简单描述，只要单击每一记录最上面一行文件名字，就可进入全文。用户既可选择某一项进行目录检索，也可以在检索框输入关键词进行检索。要想得到比较满意的检索结果，Infoseek 有一定的检索规则与算符需要遵循。

2）Infoseek 字段检索

字段检索必须遵守一定的语法规则：字段名必须小写，字段名后紧跟冒号，冒号与检索词之间不允许有空格，检索词只能是一个单词、一个短语或一个名称。

（1）单词检索与词组检索。在检索框中输入与主题相关的一个或多个单词，单击 Search 按钮便完成了一次检索。为了提高检全率，Infoseek 支持同义词检索，可以在检索框中同时输入几个含义相近的单词进行检索。如输入 restaurant（饭馆）、cafe（餐馆）、

Destinations in China	
Beijing	Macau
Chengdu	Nanjing
Chongqing	Ningbo
Dali	Qingdao
Dalian	Shanghai
Dongguan	Shenyang
Fuzhou	Shenzhen
Guangzhou	Suzhou
Guilin	Taiwan
Hangzhou	Tianjin
Hong Kong	Wuhan
Huangshan	Xiamen
Kunming	Xi'an
Lhasa	*MORE*

Provinces in China	
Anhui	Jiangxi
Beijing	Jilin
Chongqing	Liaoning
Fujian	Macau
Gansu	Ningxia
Guangdong	Qinghai
Guangxi	Shaanxi
Guizhou	Shandong
Hainan	Shanghai
Hebei	Shanxi
Heilongjiang	Sichuan
Henan	Tianjin
Hong Kong	Tibet
Hubei	Xinjiang

InfoseekChina Sites

Top Stories　China Technology News
China Entertainment News　China Beverage News
China Travel & Tourism News　China Sports News
The Tales of Grasshopper

News Media in China | Top Stories

Business　Regional
Entertainment　Sports
National　Technology
Blogs & Reports

Key Industries in China | News

Aerospace/Defense | News　Industrial Goods | News
Agriculture | News　Insurance | News
Automotive | News　Int'l Bus & Trade | News
Banking | News　Marketing | News
Biotech & Pharma | News　Metals & Mining | News
Business Services　Property Development | News
Chemicals | News　Retail | News
Conglomerates | News　Semiconductors | News
Consumer Goods | News　State-Owned Enterprises
Construction | News　Technology | News
Energy, Oil & Gas | News　Telecommunications | News
Food & Beverages | News　Textiles | News
Health & Wellness | News　Transportation | News
Hospitality | News　Utilities | News

Investing in China | News

Exchanges & Investing　Shenzhen Exchange
Hong Kong Exchange　Trade & Investment
Shanghai Exchange

图 11-62　InfoseekChina 搜索引擎主界面分类目录

bistro(小餐馆、小酒店),从而在一定程度上避免了漏检。如果要查找必须含有某词组的网页,有两种方法可供选择。一种短语需用双引号("")括起。例如"world wide web",若不用双引号,Infoseek 将查找含有 world、wide 和 web 三个单词的网页,检索结果相去甚

远。另一种可以用大写字母形式输入词组，如 WorldWideWeb，系统查找 World、Wide、Web 三个单词必须紧挨在一起的网页。

（2）短语检索与名称检索。Phrase：短语检索，即按一定次序出现的词串。短语检索词形式与多个单词组合检索词形式的区别在于短语必须用双引号括起。如“yellowbrickroad”，返回结果中将包含原检索词，并保持固有词序。否则 Infoseek 将被视为多个单词的组合，返回结果中可能包含 yellow、brick、road 中的一个或几个单词，且不一定保持原词序。与普通检索不同的是，高级检索中的短语无须用引号括起。Name：人名、公司等名称检索，高级检索中的名称可以不采用大写。word(s)：单词查询即选择一定的检索词形式后，便可在其后的空白框内输入相应的检索词。

Infoseek 的普通检索支持名称检索，包括人名和事物名称，它们必须以大写字母开头，如 SharonStone。如有两个或两个以上的名称同时作为一个检索词，则需要用逗号将它们分隔，否则将被视为一个短语。如 WhiteHouse，BillClinton。

（3）AND、OR 与 NOT 算符应用。AND 即逻辑与运算，要求查找的网页必须含有某些关键词，如检索结果中必须出现某词，在此词前标上“＋”，如 cityguide＋SanFran－cisco。例如输入：＋“ troutfishing”＋tackle equipment，检索结果必须包含 trout fishing 和 tackle，而 equipment 可有可无。增强了检索的专指性，缩小了检索范围，提高了信息的查准率。需要注意的是加号“＋” 与其后面的关键词不能留有空格。

OR 即逻辑或运算，用空格或逗号把关键词分开，表示查找的网页不必同时包含这些关键词，而只要含有其中任何一个即为命中结果。如用空格表示的例子：author writer novelist。这起到了增加检索词的同义词与近义词，扩大检索范围的作用，提高了查全率。

NOT 即逻辑非运算，如检索结果中排除某词，在此词前标上“-”，如 Python-Monty。输入“ small dog’-chihuahua，查找 small dogs（小狗），但排除 Chihuahua（一种产于墨西哥的吉娃娃狗）的网页。

（4）大小写敏感。查询的关键词，若用其小写形式，表示任何形式都匹配。如输入 california，含有 california、California 和 CALIFORNIA 的网页都会出现在检索结果中。但用大写形式 California，则只能查出含有 California 的网页。

（5）管道符检索。为了提高检准率，Infoseek 在相邻两词间使用管道符“| ”，表明对第二词的检索只在第一词的检索结果范围内进行，比如 dogs|daluations。比如 dance|tango，表示在 dance(舞蹈)这一上位类目下检索有关 tango(探戈)的信息，得到约 40 万条结果信息，比单纯输入 dance 检索得到 500 万条结果减少了不少无用信息，在一定程度上降低了误检。另一种方法是在检索结果页上，选择 Search within Result(在检索结果中查

某内容)框，输入关键词，同样可以进一步缩小检索范围。此外，Infoseek 还允许在检索框中输入多个单词来描述检索课题。如：best pizza in SanFrancisco，这样得到的结果较之单个语词的检索，其准确性得到大大提高。

(6) 标题检索。在“title：”后输入检索词，此检索词可以是单元词也可以是用双引号(“”)括起来的短语，“title:”返回网页文档标题中包含该检索词的信息。如 title：usedcar。输入 title：stamp collecting，查找网页标题名含有 stamp collecting(集邮)的文档。

(7) 网站检索。在“site:”后面输入网站域名作为检索词，“site:”返回特定站点下的网页。但如果用户想搜索某一网站上的某些信息，在“：”与后面的网站域名检索词之间不能有空格，检索词前用“＋”号，如＋site：travel city com＋Miami。例如需要检索美国广播公司(ABCnews. com) 网站上有关南非方面的文档，检索式是 site：abcnews. com＋SouthAfrica。

(8) 网址检索。在“url：”后输入一个 URL 名称。输入 url：travel，将查找网址中含有 travel 的网页。“url:”返回网页的 URL 中包含该检索词。

(9) 超文本链接检索。与某站点链接的页面检索，在/“Link：” 后输入要查与此链接的 Web 站点名，如 Link：yahoo. com。用于了解某个网站被其他网页链接的数量，“link:”返回的网页必须有包含其后检索词的链接。如输入＋link：widgets. com-site：widgets. com，查找除自己网页内部链接以外的所有链接到 widgets. com 公司的网页，以了解该公司网站受欢迎的程度。

(10) 其他信息查询。图像的查询(imageseek：)，在“imageseek:”后输入要查图像名称；在网页的文档中查找(Document：)，在“Document:”后输入要查文档名称。

11.5 雅虎搜索引擎信息检索应用

1. 雅虎概述

1994 年华人杨致远和大卫・费罗在美国于 1994 年创立了雅虎。雅虎(Yahoo!)是美国著名的互联网门户网站，也是 20 世纪末互联网奇迹的创造者之一。其服务包括搜索引擎、电邮、新闻等，业务遍及 24 个国家和地区，为全球超过 5 亿的独立用户提供多元化的网络服务，同时也是一家全球性的因特网通信、商贸及媒体公司。雅虎是最老的“分类目录”搜索数据库，也是最重要的搜索服务网站之一，在全部互联网搜索应用中所占份额较大。所收录的网站全部被人工编辑按照类目分类，其数据库中的注册网站无论是在形式上还是内容上质量都非常高。新一代雅虎搜索引擎的首页采用搜索引擎一贯的简洁风

格，以雅虎搜索的搜索框为主体，集中突出地体现出搜索的概念。见图 11-63。



图 11-63　Yahoo!搜索引擎主界面

2. 雅虎搜索引擎的搜索技术

美国雅虎最早以人工分类和网址收集见长，特别是随后斥 26 亿美元收购了可以与 Google 匹敌的 Inktomi、Overtune（全球最大的搜索广告商务提供商），Fast、AltaVista、Kelkoo（欧洲第一大竞价网站）五家国际知名搜索服务商后，经过近一年的消化和二次开发，雅虎在整合众多核心技术的基础上推出了 YST 技术。雅虎搜索引擎技术（Yahoo! search engine technology，YST）是一套基于算法的 Web 索引抓取程序，能够自动探测网络内容。YST 这套机器搜索程序从因特网上采集文档，建立起一个可搜索的索引系统。这些文件（即用户的网站文件）能被 YST 程序发现和抓取的主要原因是，在因特网其他的网页上包含有这些文档的直接链接。YST 搜索程序严格遵守 robots. txt 标准执行抓取。因此，对于那些您不希望被雅虎搜索引擎返回的结果，搜索程序不会执行抓取。任何被 robots. txt 标准认为不适宜抓取的文件，既不会被包括在抓取文档中，也不会进入到搜索引擎的数据库。目前，YST 已经成为国际两大顶级网页搜索引擎之一，也是全球使用量最高的网页搜索引擎之一。

3. 雅虎搜索引擎的基本搜索功能

雅虎网页搜索界面简洁明朗，使用方法也非常简单，输入想要查找的关键字，单击“雅虎搜索”即可。雅虎默认的设置是搜索英文结果的网页。所要检索的关键字可以是词语，也可以是短语或句子。但应注意的是，如果以短语或句子作为关键词，则必须在两端添加英文输入法状态下的双引号，否则雅虎将把短语或句子视为若干独立词语，从而同时搜索包含这几个词语的网页。例如以“网络视频会议”为关键词进行检索：“network video conference”，若加了英文双引号，Yahoo 将搜索所有包含“network video conference”整句的网页；若不加引号，则雅虎将搜索含有“network”、“video”、“conference”、“network video”、“video conference”等词语或词组的网页。见图 11-64。

雅虎搜索引擎不区分英文字母大小写，输入“yahoo”和“YAHOO”，所得结果都是一样的。用户若需要查找特定语言的网页，只需要在高级搜索的“按语言搜索”中自行设定即可。雅虎目前支持搜索用英文、中文、法文、德文、俄文、韩文等 40 种语言。

使用雅虎进行搜索，多数的搜索结果都会包含网址链接、文摘、网页快照和类似网页四项。雅虎的文摘不是通常的那种网站简介，而是对网页中那些与关键字最为相关的内

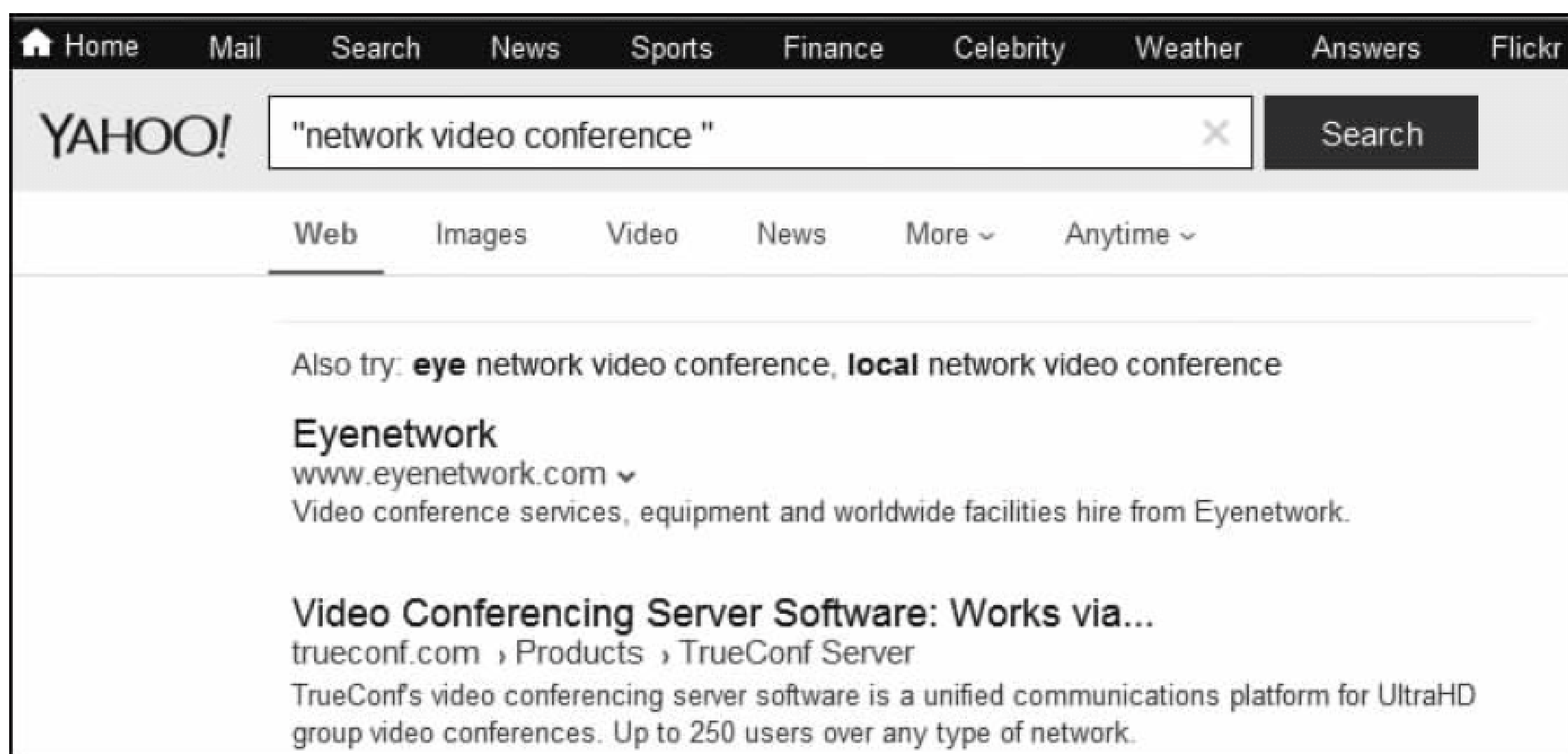


图 11-64　yahoo 检索词语加双引号后的检索实例

容的摘录;网页快照则是雅虎对它所访问过的网页的备份,这项功能使得用户在存有网页的服务器出现故障时仍可浏览该网页的大致内容;类似网页则是与当前网页内容相关的其他网页,方便用户进行对比和参考。一次搜索往往会得到数量庞大的结果,如何在这些结果中为用户选出那些最大价值的项目是每个搜索引擎首先要考虑的问题。与 Google 和百度不同的是,Yahoo 在搜索结果页面的右边设置了“按照时间(一周内、一月内、三月内)”和“按照格式(word、ppt、PDF 以及专业文档) ”两种筛选结果,便于用户能按照时间、格式迅速查找。如 PDF 文档在国际上被作为标准格式普遍使用,一般而言,网络上以 PDF 格式存储的信息往往内容比较正式,价值也相对高一些。直接单击搜索结果页面的右边的 PDF 按钮,就会出现所需资料的所有 PDF 文档,雅虎会在其标题前冠以深色的“[PDF]”字样。

4. 雅虎搜索引擎的搜索常识与技巧

如果用户已知要查找内容的主题概念,就可以利用关键词检索方式,在检索框中输入要找的关键词,然后单击“搜索”按钮,雅虎就会在数据库中查找与关键词匹配的记录,并将符合检索条件的结果显示出来。使用关键词检索还有简单方法与复杂方法之分:简单方法就是将关键词直接输入检索文本框中,可以输入一个词,也可以输入几个词,并对检索要求不加限制,系统在处理时会按照自身的规则将用户的查询字符串分为几个部分,这样返回的结果可能与用户想要的信息相差甚远;而复杂方法(或高级检索)就是利用字段限定符号和限制选项构造复杂的检索表达式来进行检索,这样会获得比较准确的查询结果。雅虎支持以下几种限定检索操作符。

（1）用引号（“ ”）来查询完全符合关键词字符串的网页。

（2）在关键词前加“ t：”，搜索引擎仅限在网站名称中查找网页。

（3）在关键词前加“ u：”，搜索引擎仅限在 URL 中查找网页。

（4）在关键词前加“＋”，查询结果中一定要出现“＋”号后面的字符串。

（5）在关键词前加“ － ”，查询结果中一定不能出现“ － ”号后面的字符串。

（6）雅虎搜索的默认的设置是包含用户输入的所有关键字。包含关键字：要在加入的词前输入一个空格。例如，用户要搜索 Paul Grein 的歌曲，可以在“Paul Grein”后面输入一个空格，再输入“music”，就能得到有 Paul Grein 歌曲的网站。

（7）在要加入的词前输入半角的加号“＋”。如果用户要搜索 Paul Grein 的歌曲，可以输入 Paul Grein + music，出现的搜索结果就是带有 Paul Grein 歌曲的网站。

（8）去除关键字：与包含关键字相反，想要去除一个关键字，用户需要在这个词前输入减号“ － ”，但在减号之前必须留一个空格。例如，用户想要找除了摇滚以外的音乐信息，只要在搜索框里输入“music Rock”（ 注意，music 后要加空格）即可。

（9）尝试使用特定的搜索词汇去描述要找的内容。通常，比较广义的关键字搜索出来的结果会很多，而当用户想要更精确的搜索结果时，最好选用一些狭义的关键字做搜索。如用“digital camera”取代“camera”。

本 章 小 结

作为新时代的大学生，应用搜索引擎去充分发现、认识、查询、获取和有效利用网络信息，不仅是大学生信息检索素养的重要组成部分，也是开展自主学习、协同学习、探究性与研究性学习的基础性信息素养及其内在要求。搜索引擎（search engine）是一种网络化信息检索系统与检索应用工具，能帮助用户在浩瀚的网络资源环境中快速而高效地查询到所需要的信息。搜索引擎是一种能够通过网络接收用户的查询指令，并向用户提供符合其查询要求的信息资源网址或资源路径的智能系统。

作为普通用户而言，经常接触到的是网络搜索引擎的用户检索交互界面。用户检索交互界面是搜索引擎各种检索实现功能在用户接口层面直接而形象的表达，屏蔽了搜索引擎所应用的各种检索原理、检索技术与数学逻辑过程。用户检索交互界面的作用是接收用户的各种查询输入、显示查询结果、提供相关反馈信息。用户检索界面包括简单检索界面和高级检索界面两类。简单检索界面只提供用户输入查询字符串的文本框，高级检索界面提供用户按照各类检索模型的查询机制（例如查询范围限制、信息筛选与过滤、多

个检索词之间的逻辑组配等)。

本章详细阐述了大学生常用的搜索引擎,包括百度、搜狗、谷歌、雅虎和 Infoseek,重点说明了它们的多种搜索服务功能与主要信息查询方法。第一,用户要了解和熟悉搜索引擎的各种服务产品,例如百度搜索有百度学术、网页、视频、音乐、新闻、图片、软件等二十多种服务产品。第二,大学生用户要逐步养成应用高级检索功能和搜索个性设置的良好检索习惯,以利于提高检索结果的效率与准确性。第三,不同的搜索引擎都有自身的特色,注意掌握一些常用搜索技巧,例如查询词的明确表达与简练化、检索词的精确匹配、指定网站或指定网页内搜索、特定检索语法的应用等。

本章思考与练习题

1. 什么是搜索引擎? 常用的搜索引擎有哪些?
2. 请举例说明你使用搜索引擎有哪些方法。
3. 百度引擎有哪些核心技术?
4. 百度引擎有哪些主要信息搜索服务产品?
5. 百度引擎移动搜索端与 PC 搜索端有差异吗? 请举例说明。
6. 百度网页搜索有哪些主要方式? 分别举例说明。
7. 百度引擎的常用检索技巧有哪些? 分别举例说明。
8. 搜狗搜索入门应该从哪几个方面着手?
9. 不能正常访问搜狗引擎有哪些常见的解决办法?
10. 使用搜狗搜索有哪些主要技巧?
11. 搜狗搜索与百度引擎的高级检索有差异吗? 请举例说明。
12. 搜狗信息搜索服务有哪些主要服务产品?
13. Google 与 Baidu 搜索引擎的翻译功能有差异吗? 翻译准确度方面有差异吗? 请举例说明。
14. 举例说明快捷有效的 Google 特殊操作符的搜索应用如何。
15. Google 信息检索有哪些实用功能?
16. Google 高级搜索有哪些主要功能? 请举例说明。
17. Infoseek 有哪些主要的检索应用方法? 请举例说明。
18. 雅虎搜索引擎有哪些基本的搜索功能? 请举例说明。
19. 使用雅虎搜索引擎应该注意哪些搜索常识与技巧? 请举例说明。

第 12 章　特种信息资源检索

对于大学生或科技工作者而言，特种信息资源是指出版发行和获取途径都比较特殊的科技类信息资源，通常也指除了普通图书信息资源和期刊信息资源之外的特种科技信息资源。它们通常包括会议文献信息资源、科技报告信息资源、专利信息资源、学位论文信息资源、标准信息资源、科技档案信息资源、政府出版物信息资源七大类。特种信息资源特色鲜明、内容广泛、数量庞大、学习与研究及其参考价值高，在整个信息资源与信息检索及其利用过程中起着非常重要的作用。特种信息资源的载体形式丰富，除了光盘与印刷型纸质载体外，目前大多数也以网络数据库的形式提供检索服务。

12.1　科技报告信息资源检索

12.1.1　科技报告的概念与特征

1. 科技报告的概念

科技报告（scientific & technical report）是指对科学、技术研究成果或研究进展的记录，也称研究报告或报告文献。科技报告的出现早于科技期刊，在科学交流制度化之前科技工作者们就已经生成各类科技报告了。但是，作为一种传递科技信息的特定类型的信息资源，其历史能追溯到 20 世纪初。当时，只是研究者或设计单位向经费支助机构提交关于研究或设计任务完成情况以及财务支出情况的报告，大量的研究成果以内部报告交流的形式出现。

2. 科技报告的特征

（1）内容特征。一是迅速反映新的科研成果，以科技报告形式反映科研成果比这些成果在期刊上发表，一般要早一年左右，有的则不在专业期刊上发表。第二是内容多样化，科技报告几乎涉及整个科学、技术领域和社会科学以及部分人文科学领域。第三是保密性，大量科技报告都与政府的研究活动、高新技术有关，使用范围控制较严，一般只在同类性质的机构内部交流，公众难以获取。最后是真实性和专业性，科技报告反映的内容直接来自实际工作和研究，有大量的事实、数据、结论、建议等，阅读对象也主要是专业对口

的科技人员和管理人员，审查也多是专业人员和机构。

（2）形式特征。①每份报告都有统一编排的报告号，报告号通常是以研究的执行机构或主管部门的缩写字母加上顺序号组成，一般不会变更。报告号既是每份科技报告的入藏、排架号，又是提供使用、复制和订购时的索取号。②具有统一的格式和比较完整的信息标识项目。科技报告的篇幅不受限制，可长可短。少的几页，多的数千页。但不管内容多少，都有统一的编写规格，主要包括报告题名、统一封面、目次、文摘、序言、报告主体和附录等。同时报告标题、入藏号、团体著者、报告号、个人著者、任务号、合同号等均加以数据标引。③具有冗长的篇名。这是科技报告不同于其他信息源的最突出的特点。图书、期刊、专利、标准等文献信息的篇名，一般只有2～5个主题词或关键词，而科技报告由于专业技术性强、内容具体，所以篇名特别长，一般有5～15个关键词。科技人员只需看篇名即可了解其大致内容。④每份报告为一项专题材料，自成一册。

12.1.2 科技报告的类型与编码

1. 科技报告的类型划分

（1）按科技报告反映的研究阶段划分：①初期报告（primary report）或开题报告，是研究机构对研究项目的一个计划性报告；②中期报告或过程报告，如研究过程中的现状报告（status report）、预备报告（preliminary report）、中期报告（interim report）、进展报告（progress report）、非正式报告（informal report）；③结题报告或总结报告，即研究工作结束时的报告，如总结报告（final report）、综述报告（definitive report）、试验结果报告（test results report）、竣工报告（completion report）、正式报告（formal report）和公开报告（public report）等。

（2）按报告的使用秘密等级划分：①秘密报告（secret report），分为绝密报告（top secret report）、机密报告（confidential report）和秘密报告（secret report）三类，供少数人员查阅；②非密/限制发行报告（unclassified/limited or restricted report），只在规定范围内发行，数量也有限定；③解密报告（declassified report），即曾经是保密的科技报告，但经过一段时间后失去保密意义，解密为公开发行的报告；④非密/解除限制发行报告（unclassified/delimited report）等。

（3）按报告的内容性质划分：有科学报告（science report）、技术报告（technical report）、工程报告（engineering report）、调查报告（investigation report）、研究报告（research report）、专门报告（special report）、分析报告（analysis report）、会议报告（conference report）、评估报告（evaluation report）、专题报告（topical report）、交流报告

(circular report)、生产报告(production report)、经济报告(economic report)等许多类型。

2. 科技报告的编码

科技报告都有一定的编号特征,但各个系统和单位的编号方法并不一致。科技报告的常见代号一般有以下几种类型。

(1) 机构代号:机构代号是科技报告编码的重要部分,一般用编辑、出版、发行机构名称的首字母,标识在报告代号的首位。

(2) 类型代号:主要代表科技报告的类型。有的用缩写字母表示,如 PR 报告(进展报告);有的用数字表示,如 DOE 报告的"TID-5000"代表研究发展报告等。

(3) 密级代号:代表科技报告的保密情况。如 ARR(绝密报告)、S(机密报告)、C(秘密报告)、R(控制发行报告)、U(非保密报告)等。

(4) 分类代号:用字母或数字表示报告的主题分类,如 P——物理学(Physics)等。

(5) 日期代号和序号:用数字表示报告出版发行年份或报告的顺序号,例如 STAN—CS—92—920,即 STAN—CS(机构)—92(年份)—920(序号)等。

12.1.3　国内科技报告与商业报告资源的信息检索

1. 国家科技成果网

(1) 概述。国家科技成果网(http://www.tech110.net/portal.php)是由中华人民共和国科技部科技成果管理办公室和中国化工信息中心承办的一个全国性科技成果信息服务平台,主要设置了成果查询、成果登记、成果公报、成果统计分析、网上成果展等信息资源内容。

(2)成果的简单检索。网站包括了国内各个科技领域的重要成果。首先在主界面的菜单条上单击"成果",然后在输入框中输入要搜索的成果关键词,单击"搜索"按钮后就可实现科技成果的简单检索。见图 12-1。



图 12-1　国家科技成果网一般检索用户界面

(3) 分类目录检索。如果不确定自己详细的检索需求和不能拟定准确的检索词,可以依据其详细的一级、二级和三级目录进行细化。其中一级目录有六大类:“农业·化工”、“生物·医药”、“能源·采矿”、“建筑·水利”、“交通·运输”和“自然·社科”。

① 假设用户需求以“交通·运输”为例,其详细的一级、二级和三级目录如图 12-2 所示。

图 12-2　国家科技成果网的分类目录检索实例

② 通过图 12-2,选择三级目录下的“无线电设备、电信设备”后,返回检索结果 2283 项。为了便于用户对比、细化筛选条件和检索范围,国家科技成果网进行了详细的成果分类统计,实例如图 12-3 所示。

单击“高级搜索”按钮,即可进入高级检索界面。在这个页面中可以设置更灵活的搜索条件,来完成更复杂的查询。

检索结果以列表的方式显示成果名称,每页只能显示 20 条。单击成果名称,可浏览成果详细信息。免费会员仅能看到成果题录信息和文摘,如要看全文,需交纳相应费用成为付费会员才可看到。

(4)高级检索。国家科技成果网的高级检索主要提供关键词或主题词的与(AND)、或(OR)、非(NOT)三种布尔逻辑检索以及检索项的查询筛选控制。检索项的查询筛选控制条件包括成果类别、关键词、技术成熟度、成果简介、应用行业、登记日期、课题来源等近 20 项内容,高级检索界面实例如图 12-4 所示。

2. 万方中文科技报告数据库

我国研究成果的统一登记和报道工作是从 1963 年正式开始的。凡是有科研成果的

"无线电设备、电信设备(TN8)" 相关 成 果 共找到 2283个结果　已选条件：		
成果类别	应用技术(2254)　软科学(2)　基础理论(27)	
单位所在省市	广东省(319)　江苏省(240)　上海市(201)　北京市(195)　浙江省(164)　陕西省(151)　四川省(124)　天津市(106)　安徽省(104)	更多
课题来源	国家科技计划(176)　自选课题(430)　民间基金(3)　横向委托(38)　国际合作(1)　地方基金(47)　部门基金(22)　地方计划(180)	
所属高新技术类别	电子信息(1250)　先进制造(192)　航空航天(16)　现代交通(0)　生物医药与医疗器械(1)　新材料(69)　新能源与节能(45)	更多
应用状态	产业化应用(913)　小批量或小范围应用(24)　试用(13)　应用后停用(4)　未应用(62)	
推广形式	其他(524)　技术服务(172)　合作开发(119)　产权转让(62)　技术入股(53)　资金入股(48)	
成果发布年份	2007(284)　2002(205)　2006(190)　2001(161)　2009(150)　2004(140)　2012(139)　2003(137)　2010(131)　2011(117)	更多
研究形式	独立研究(856)　与企业合作(66)　与院校或院所合作(0)　与国外合作(0)　其他(26)	
成果体现形式	新技术(302)　新工艺(0)　新产品(476)　新材料(28)　农业、生物新品种(0)　矿产新品种(0)　新装备(30)　其他应用技术(0)	更多
应用行业	农、林、牧、渔业(2)　采矿业(0)　制造业(1870)　电力、热力、燃气及水的生产和供应业(48)　建筑业(2)　批发和零售业(3)	更多
技术成熟度	初期阶段(94)　中期阶段(237)　成熟应用阶段(687)	
单位属性	独立科研机构(97)　大专院校(156)　企业(709)　医疗机构(4)　其他(49)	
成果完成人	康洪(11)　吴德君(11)　张涛(10)　王伟(10)　王勇(9)　王文(9)　李树林(9)　王跃(8)　杨维华(8)　张勇(8)　熊绍珍(8)	更多
技术水平	国际领先(99)　国际先进(321)　国内领先(540)　国内先进(181)　国内一般(109)	
中图分类	显示设备、显示器(427)　电源(386)　光纤传输线、光缆(230)　天线(75)　其他(44)　测试、调整及设备(42)	更多
成果登记日期	1997-10-31(66)　1998-10-31(26)　1995-10-31(23)　1996-10-31(22)　1993-10-31(22)　1994-10-31(20)　1999-10-31(11)	更多

图 12-3　三级目录"无线电设备、电信设备"的检索结果实例

首页 > 搜索中心 > 高级搜索

成果　专家　机构

高级搜索

逻辑运算优先级如下：NOT > AND > OR

于　所有字段和文本

AND　于

AND　于

添加行 | 删除行

所有字段和文本
所有字段(无全文文献)
更多选项
成果类别
关键词
成果体现形式（应用技术类）
技术成熟度
技术水平
研究形式
中图分类号
所属高新技术类别
应用行业
课题来源
应用状态
推广形式
成果简介
成果登记日期
单位属性
单位所在省市
成果完成人
成果发布年份

搜索选项

成果登记日期：　至　[格式：YYY

成果发布年份：　至　[格式：YYY

排序依据：　相关性

每页显示条目数：　20　搜索　清空搜索选项

Copyright 2001-2020 All Rights Reserved© 国科网

国家科技成果信息服务平台 主管单位：国家科学技术奖励工作办公室

图 12-4　国家科技成果网高级检索界面

单位都要按照规定程序上报、登记，1971 年起统一定名为《科学技术研究成果报告》。

检索我国科技成果报告可通过万方数据资源系统中的《中文科技报告数据库》。该库始建于 1986 年，收录了自 1966 年至今的历年各省、市部委鉴定后上报国家科技部的科技成果报告，共 40 万余条科技成果。可供公共查询的是经过中华人民共和国科学技术部审批并已公开的中文科技报告 20 000 余份，专业涉及化工、生物、医药、机械、电子、农林、能源、轻纺、建筑、交通、矿冶等诸多领域。这些领域分成八大部分：国家重大科技专项、国家重点基础研究发展计划、国家高技术研究发展计划、国家科技支撑计划、国家国际科技合作专项、国家重大科学仪器设备开发专项、国家科学技术奖励项目、国家重大科学研究计划。

作为各个高校数字化校园的服务资源之一，一般都购买了万方数据资源系统，大学生通过本学校的局域网可以免费检索；如果所在学校没有购买，用户需要通过互联网检索并预先付费，获得用户名和密码后才可进行检索，检索网址为：http://c.wanfangdata.com.cn/NSTR.aspx，其检索分类目录如图 12-5 所示。



图 12-5　万方中文科技报告分类检索目录

检索项包括成果名称、成果题名、作者、关键词和成果完成的起止时间。例如，通过分类一级目录“国家科技重大专项”选择其中的二级目录“新一代宽带无线移动通信”获得相

关科技报告 163 条,实例如图 12-6 所示。



图 12-6　万方中文科技报告数据库检索实例

3. 国务院发展研究中心报告(国研报告)

国务院发展研究中心是直属国务院的政策研究和咨询机构,是国内国际知名的政策研究咨询机构,在宏观经济政策、发展战略和区域经济政策、产业经济和产业政策、金融以及国际经济等领域拥有许多国内外著名的经济学家以及高素质的专家和研究人员。国务院发展研究中心信息网(简称"国研网")由国务院发展研究中心主管、北京国研网信息有限公司承办,创建于 1998 年 3 月,是中国著名的专业性经济信息服务平台。

进入国研网主页,在检索输入框中输入关键词,如果有多个关键词,关键词间可以使用逻辑算符连接。在该检索系统中,表示"且"的关系,使用空格、"+"或"&";表示"非"的关系,使用字符"—";表示"或"的关系,使用字符"|";如果表达式是一个整体单元,使用字符"()"。单击"检索"按钮,系统显示题名与摘要。选择需要查看全文的报告,单击"标题名称"就可以看到报告的全文。一般文科或综合性高校都购买了国研报告数据库,检索和阅读全文是免费的。如果某些类型的高校(例如工科类高校)没有购买,则用户需要网络注册后使用。国研报告的分类检索目录见图 12-7。

国研报告的分类检索目录依据经济产业的行业进行分类,便于分类查询。检索方式有关键词、标题、作者与全文。假设依据其"月报"为例,获得的检索结果如图 12-8 所示。

4. 中国商业报告数据库

(1) 普通检索。中国商业报告数据库(http://www.chinainfobank.cn)是中国资讯行的子库之一,收录经济专家及学者关于中国宏观经济、金融、市场、行业等的分析研究文献及政府部门颁布的各项年度报告全文,主要为用户的商业研究提供专家意见的资讯,数

国研网系列研究报告　　最新文章 | 产品名录

· 周评　· 国内金融形势与货币政策周评　· 银行业监管与市场动态周评　· 信托业监管与动态周评　· 证券业监管与动态周评　· 期货业监管与动态周评　· 保险业监管与动态周评　· 国际金融形势周评

· 月报　· 宏观经济　· 房地产业　· 电力行业　· 钢铁行业　· 汽车行业　· 石油化工　· 金融中国　· 通信行业　· 港口航运　· 医药行业　· 电子商务

· 季报　· 宏观经济　· 房地产业　· 电力行业　· 石油行业　· 化工行业　· 钢铁行业　· 汽车行业　· 金融中国　· 交通运输　· 医药行业　· 生物制药　· 通信设备　· 通信行业　· 食品制造　· 电子商务　· 水泥制造

· 年报 New!　· 房地产业　· 电力行业　· 钢铁行业　· 汽车行业　· 化工行业　· 医药制造　· 食品制造　· 通信行业

图 12-7　国研报告的分类检索目录

国研网系列研究报告

月　报

宏观经济 / 金融中国 / 房地产业 / 电力行业 / 钢铁行业 / 汽车行业 / 石油化工 / 通信行业 / 港口航运 / 医药行业

您当前位置：首页 > 国研网系列研究报告 > 月报

宏观经济　>> 更多

· [国研专稿]2016年4月运行指标分析　2016-05-19
· [国研专稿]中国经济一季度企稳，仍应继续推进结构性改革　2016-05-17
· [国研专稿]央行一季度货币政策报告解读　2016-05-16
· [国研专稿]PPP模式的国际经验与我国的实施路径-现状篇（上）　2016-05-09
· [国研专稿]PPP模式的国际经验与我国的实施路径-现状篇（下）　2016-05-09

金融中国　>> 更多

· [国研专稿]2016年4月货币运行分析　2016-05-16
· [国研专稿]2016年3月货币市场运行分析　2016-05-10
· [国研专稿]2016年4月期货市场运行分析　2016-05-06
· [国研专稿]2016年4月外汇市场运行分析　2016-05-05
· [国研专稿]2016年4月份期货业监管与市场动态(下)　2016-05-05

房地产业　[国研专稿]2016年4月份期货业监管与市场动态(下)　>> 更多

· 2016年5月（上半月）房地产行业要闻综述　2016-05-16
· 2016年4月（下半月）房地产行业要闻综述　2016-05-03
· [国研专稿]2016年1-3月房地产市场运行指标分析　2016-04-29

图 12-8　国研报告的"月报"检索实例

据库每日更新。首先用户进行数据库选择，也可以默认搜索全部数据库。检索数据库包括中国经济新闻库、中国商业报告库、中国法律法规库等 14 个检索库。一般检索界面提供的检索功能有以下几项：

① 库选择，对 14 种数据库进行选择。

② 时间选择与过滤：前一月、前二月、前三月、前一年或全部时间范围五种。

③ 检索范围限定：检索词的标题位置或全部内容范围。

④ 检索词逻辑关系：全部字词出现、任意字词出现或全部字词不出现。

⑤ 检索词：任意字词，任意标题词、关键词或主题词。

(2) 检索结果的继续过滤查询。图 12-9 是以“大学生就业”为检索词在中国商业报告库中的检索结果共 41 条,依据时间倒序排列的示例图。为了用户继续评价与评定检索的返回结果,数据库还提供了“重新检索”、“同一检索命令在其他库中检索”、“在前次结果中检索(即二次检索)”等功能。

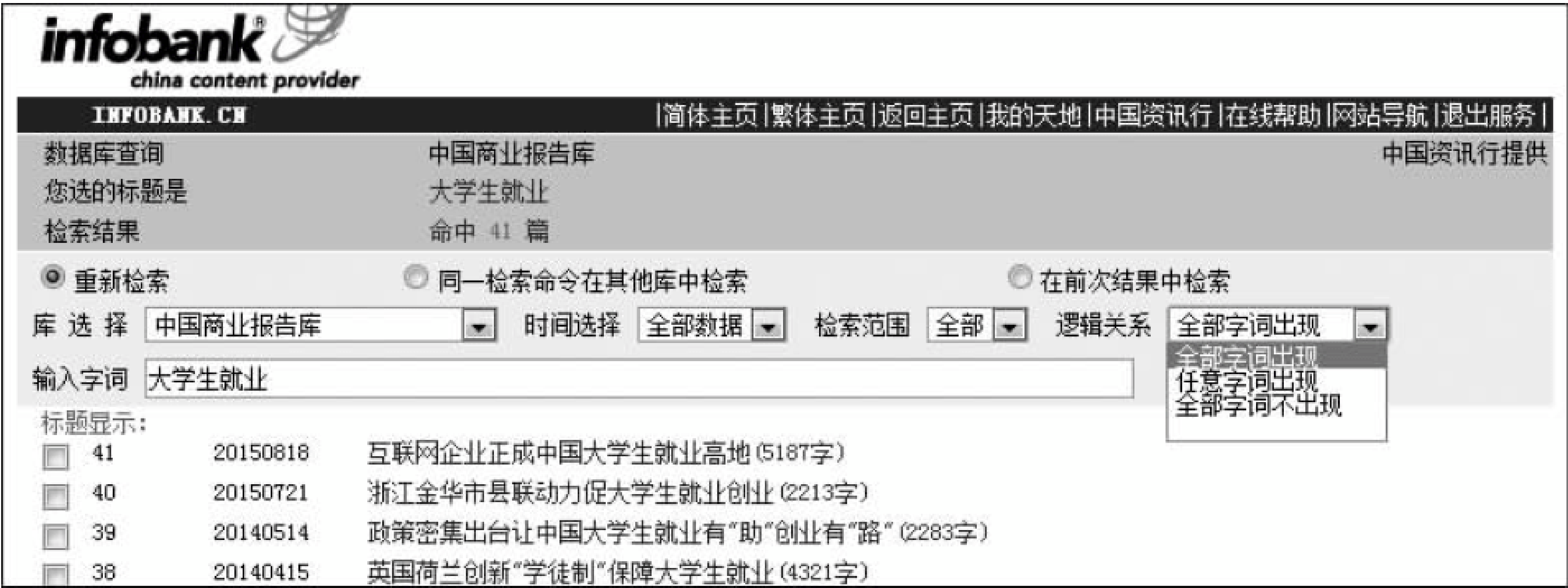


图 12-9　中国商业报告数据库一般检索实例

(3) 专业检索。中国商业报告数据库专业检索内容包括如下。

① 对行业分类的限定:默认为全部行业,也可在国防、人口、测绘、教育等各行各业中选择其一。

② 地区分类:相对来说,该分类是本数据库最详细的,包括我国省市、经济区域和世界各个国家。

③ 报告的文献出处:这部分的信息过滤主要是过滤报告的信息来源,包括很多研究所(例如国家经贸委经济研究中心报告)和丰富的经济类学术刊物(例如财经研究、东方经济等数十种刊物)。

④ 逻辑关系:全部字词出现、任意字词出现或全部字词不出现。

⑤ 检索范围:标题、副标题、正文或全部。

⑥ 返回记录数:20、50 和 100。

中国商业报告数据库专业检索界面见图 12-10。

12.1.4　国外科技报告资源检索

世界上许多国家都有科技报告的生产和收藏。比较重要的有美国四大科技报告;英国原子能管理局的 UKAEA 报告、科学与工业研究部的 DSTR 报告、航空研究委员会的 BARC 报告;日本东京大学原子核研究所报告、三菱技术通报、科学技术厅航空宇宙技术



图 12-10　中国商业报告数据库专业检索界面

研究所的 NAL-TM 报告；法国原子能委员会的 CEA 报告；加拿大原子能有限公司的 AECL 报告；德国航空研究所的 DVR 报告以及苏联的科学技术总结等。

在世界各国数量庞大的各类科技报告中，美国拥有的比重约占世界上出版的所有科技报告的 50%以上，而且比较系统化。其中，历史悠久、报告量多、参考和利用价值大的主要有 PB、AD、NASA 和 DE 报告。美国四大报告的累积量达 100 万篇以上，其中我国万方(http://c.wanfangdata.com.cn/NSTR.aspx)知识服务平台收录并提供网络服务的美国四大报告达 110 多万篇(1958 年至今)。见图 12-11。



图 12-11　我国万方数据提供的美国四大报告检索服务实例

1. PB 报告

1945 年 6 月，美国成立商务部出版局(U. S Department of Commerce Office of the Publication Board)，负责整理并公布从第二次世界大战战败国获取的科技资料，并编号出

版，号码前统一冠以 PB 字样。20 世纪 40 年代的 PB 报告(10 万号以前)主要为战败国的科技资料，50 年代起主要是美国政府科研机构及其有关合同机构的科技报告。PB 报告的内容绝大部分属科技领域，包括基础理论、生产技术、工艺、材料等。20 世纪 80 年代后，PB 报告统一采用“PB＋年代＋顺序号”的形式，如 PB97—127864。我国万方数据提供的 PB 报告达 29 万多篇，其检索项有报告提名、作者、关键词和起止时间共四项，实例如图 12-12 所示。

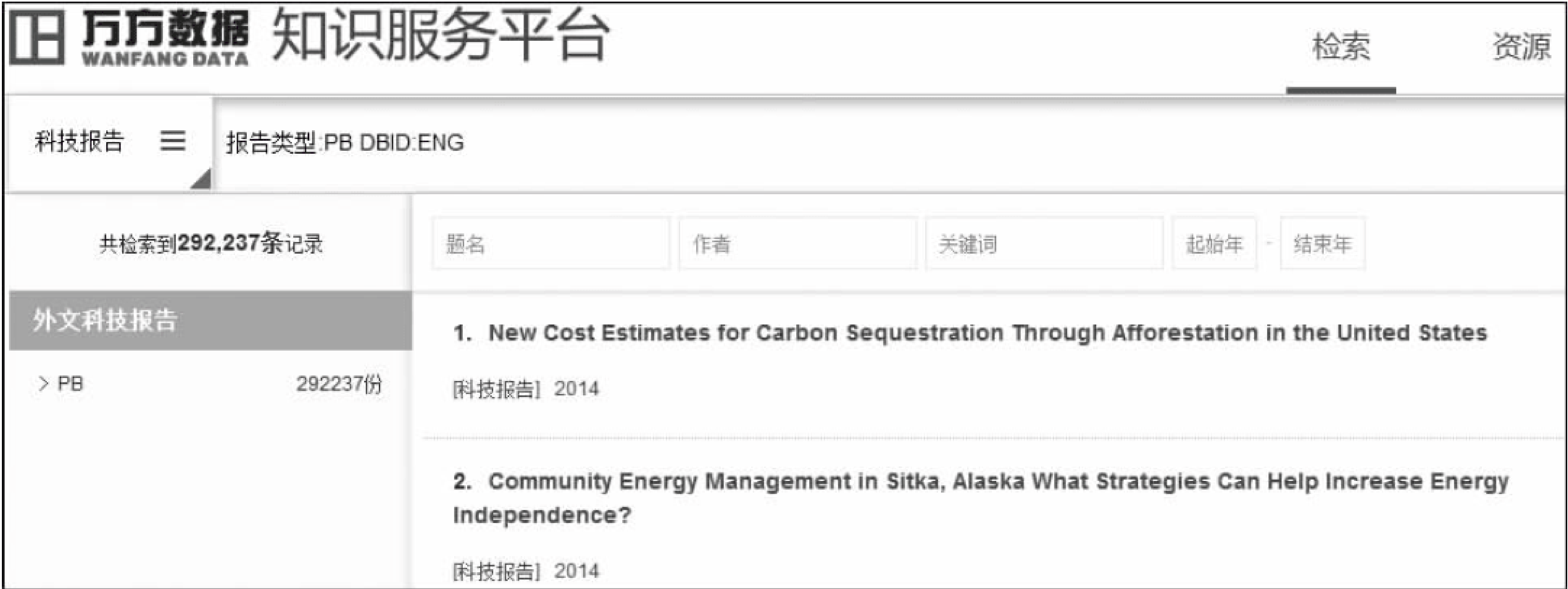


图 12-12　万方 PB 检索实例

2. AD 报告

1951 年 5 月，美国成立武装部队技术情报局(Armed Service Technical Information Agency，ASTIA)，负责收集、整理、编辑、出版国防部所属海陆空三军军事系统科研机构及其与国防部订有合同的工业、企业、高等院校提出的军事科研报告。AD 报告即是该情报局出版的文献。ASTIA 几经改组易名，但报告仍沿用 AD 名称。万方数据提供的 AD 报告达 42 万多篇，其检索项有报告提名、作者、关键词和起止时间共四项，实例如图 12-13 所示。

AD 报告有密级，并用不同的字母表示。自 1975 年起它的主要形式如下：AD—A000001～，A 表示公开报告，占 45%；AD—B000001～，B 表示非密限制报告，占 39%；AD—D000001～，D 表示美国专利文献；另外还有 AD—E 是临时使用的试验号；AD—P 是丛书或会议论文集的单行本：AD—R 是国防部和能源部能源学科的保密文献。

3. NASA 报告

NASA 报告是美国国家航空与航天局(National Aeronautics and Space Administration，NASA)拥有的研究机构产生的技术报告。该局成立于 1958 年，其前身

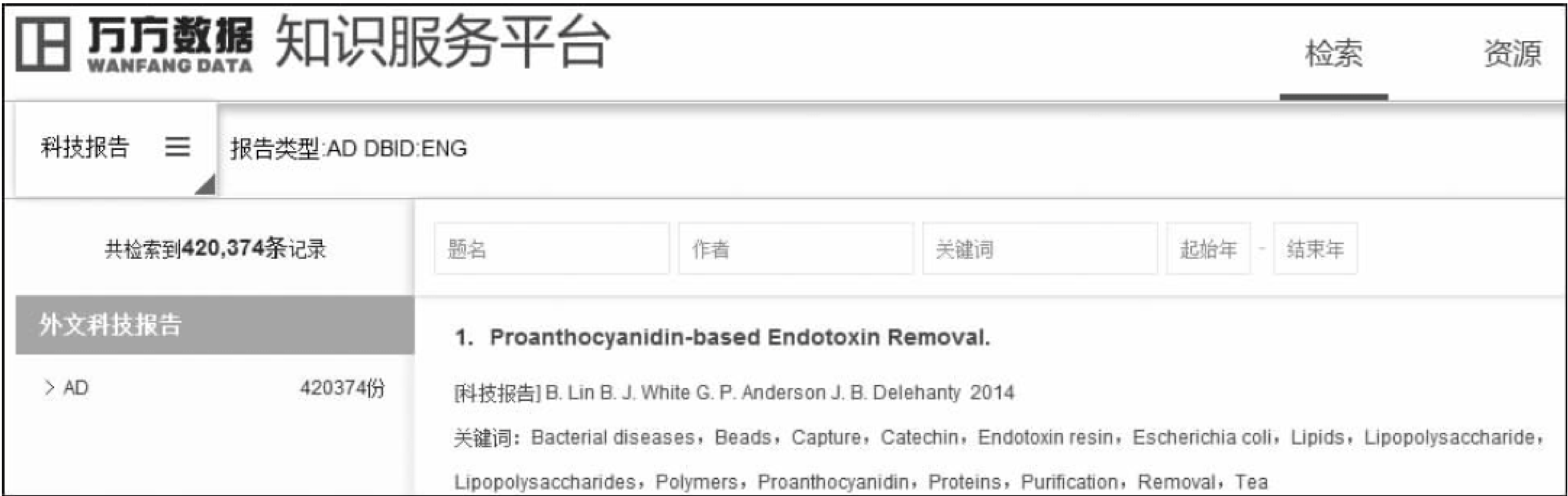


图 12-13　万方 AD 检索实例

是美国国家航空咨询委员会(National Advisory Committee for Aeronautics NACA)。NACA 报告创刊于 1915 年，主要内容为空气动力学、发动机及飞行器结构、试验设备、飞行器的制导及测量仪器等。我国万方数据提供的 NASA 报告达 11 万多篇，其检索项有报告提名、作者、关键词和起止时间共四项，实例如图 12-14 所示。



图 12-14　万方 NASA 检索实例

NASA 报告是一种综合性科技报告，除航空航天技术外，还涉及电子、机械、化工、冶金、天体物理等相关学科。NASA 报告中还包括专利文献、学位论文和专著及一些外国文献、译文等。NASA 报告号采用“NASA＋出版类型＋顺序号”的形式，如 NASA—TP—107279。报告类型主要有 NASA—TR—R—(技术报告)、NASA—TN—D—(技术札记)、NASA—TT—F—(技术译文)、NASA—SP—(特种出版物)等十余种类型。在 NASA 数据库中，NASA 文献一律冠以字母 N，其编号形式为：“N＋年代号＋顺序号”。

4. DE报告

1946年美国建立原子能委员会(Atomic Energy Commission,AEC),AEC报告即为该委员会所属单位及其合同户编写的报告。1975年,该委员会更名为能源研究与发展署(Energy Research and Department Administration,ERDA),AEC报告相应改称为ERDA报告。1977年,该署又扩大为美国能源部(US Department of Energy,DE),1978年7月起逐渐冠以DE报告,内容仍以原子能和其他能源为重点,其文献主要来自能源部所属的技术中心、实验室、信息中心和一些国外研究机构。我国万方数据提供的DE报告达31.9万多篇,其检索项有报告提名、作者、关键词和起止时间共四项,实例如图12-15所示。



图12-15 万方DE检索实例

12.2 会议文献资源检索

12.2.1 会议文献资源的概念

1. 专业会议类型

随着科学技术的发展,世界各国的学会、协会、研究机构及国际性学术组织举办的各种学术会议日益增多。

(1)按组织形式和内容,会议分为九类:Congress(专业会议),Convention(代表大会),Conference(大会),General Assembly(全体会议),Seminar(学术讨论会),Colloquium、Symposium(座谈会或学术报告会),Workshop(业务讨论会),Working Group、Discussion Group or Expert Group Meeting(工作小组、讨论小组或专家小组会议),Committee(委员会)。

(2) 按级别和范围把会议分为四类：国际性会议(包括世界各大洲都有代表参加的“世界会议”即 World Conference 和某个国际性组织或两个以上国家联合召开的“国际会议”即 International Conference)、全国会议、地区会议(一个国家的地区性学术机构单独或联合召开的)和基层会议等。

2. 会议文献的概念

会议文献(conference literature)就是指在学术会议上宣读和交流的论文、报告及其他有关资料，并且多数以会议录(proceeding)的形式出现。世界上每年产生的会议论文约10 万篇，每年出现的各种会议录就达 3000 余种。

12.2.2 会议文献的特点与类型

1. 会议文献的类型

会议文献种类繁多，出版形式多样，通常按时间把会议文献分为以下三类：

(1) 会前文献(pre-conference literature)：指在会议之前预先印发或出版的会议资料。包括会议预告(forthcoming conference)、征文启事和会议通知书，会议日程表(program)，会前论文摘要(advanced abstracts)和预印本(preprints)等。其中预印本是在会前 5～7 周内发给与会者或公开出售的会议资料，比会后正式出版的会议录要早1～2 年，但内容的完备性和准确性不及会议录。据 UNESCO 报道，约有 50%的会议只有会前文献，而不出版会议录，因此预印本显得更加重要。

(2) 会中文献：包括开幕词、讲话或报告(reports)、讨论记录、会议决议和闭幕词等。许多内容价值并不大。

(3) 会后文献(post-conference literature)：是指会议结束后，经会议主办单位等机构正式出版的会议论文集。包括会议录(proceedings)、论文汇编(transactions)、会议摘要(digest)、会议出版物(publications)等。其中，会议录是会后将论文、报告及讨论记录整理汇编而公开出版或发表的系统化文献，价值较大。

2. 会议文献的特点

(1) 内容新颖，传递及时。大多数研究先在会议上首次公布，经过一段时间才陆续在期刊或其他文献上发表，有的则根本不发表。因此，会议文献传递的是新颖的但尚未成熟的科研中的信息，远比科技期刊迅速和直接。

(2) 专业性和针对性强。科技会议都有一定的专业性，讨论的主题大都是当前人们共同关注的科学热点与难点问题，一般要邀请有关的专家学者参加，而且会议论文在会前要经过专家的评审。因此会议文献能够反映某一学科或专业的当前水平和发展动向，是

一种重要的信息源。

(3) 出版和发行方式灵活多样。通常,以期刊出版的会议录约占会议文献的 2/5;其他的会议文献或汇编成专题论文集,或出版会议丛刊、丛书,或以科技报告形式出版。有的会议文献还以录音、录像带或网络数据库形式提供服务。

12.2.3　国外会议文献的检索

1.《世界会议》

《世界会议》(World Meetings,WM)由美国世界会议信息中心(World Meetings Information Center Inc.)编辑,Macmilan Publishing Company 出版。WM 是专门预报未来两年内将要召开的世界各国学术会议信息的工具,包括国际会议、全国性会议和地区性会议,收录世界上 100 多个国家和地区的 2000 多个科技方面的专业会议情况,查询网址:http://www.wmforum.org。报道范围包括自然科学、工程技术、社会科学和医学等学科领域,由以下四个分册构成:

(1) World Meetings: United States&Canada: 1963 年创刊,预报美、加两国近两年内将要召开的各种学术会议。

(2) World Meetings: Outside United States&Canada: 1968 年创刊,专门预报美、加两国以外当年和次年将要召开的各种学术会议。

(3) World Meetings: Medicine: 1978 年创刊,报道全球两年内将要召开的医学方面的学术会议。

(4) World Meetings: Social&Behavioral Science,Education&Management: 1971 年创刊,报道全球两年内将要召开的社会学、行为科学、教育学及管理学等方面的学术会议。

WM 的四个分册都是季刊,而且编排方法和著录格式基本相同,都由正文和索引两部分组成。WM 的正文部分,即主要款目(main entry section)较详细地著录了即将召开的各种会议消息,包括会议名称、内容、召开日期和地点、主办机构及提交论文期限等。各种会议消息都会在正文部分连续报道三次,报道内容每年完全翻新一次,每期删除内容重复三次的会议,并补充最新的会议消息,从而动态地构成其报道内容的主体。

WM 的索引主要有六个,即关键词索引(keyword index)、会议日期索引(data index)、会议截稿日期索引(deadline index)、会议地点索引(location index)、出版物索引(publication index)、主办单位指南与索引(sponsor directory and index)。

2.《会议论文索引》

《会议论文索引》(Conference Papers Index,CPI)由英国剑桥科学文摘社(Cambridge Scientific Abstracts Co.)编辑出版,月刊。它主要报道世界上已经召开或即将召开的各种学术会议上宣读或递交的学术论文,报道范围涉及自然科学、工程技术和医学等领域,年报道量约10万篇。CPI作为一种题录式报道工具,既有印刷版,也有机读数据库和网络数据库,通过DIALOG、BRS或ESA/IRS系统以及CPI的机构网址http://www.proquest.com都可以进行检索。CPI检索主界面实例见图12-16。

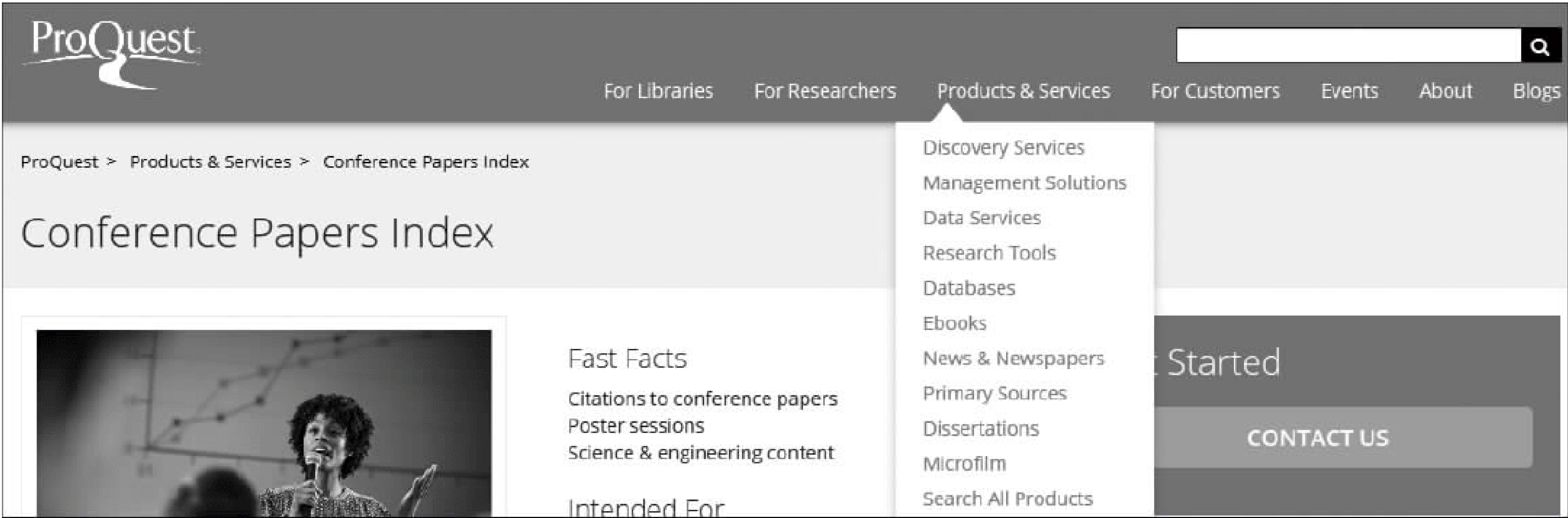


图12-16　CPI检索主界面实例

CPI由正文和索引两部分组成。正文部分是会议消息和会议论文的标题,按17个学科专业分类排列,每一类目下列出该类的各种会议的名称、召开日期及地点、订购消息等项。紧接着会议消息之后著录了会议上即将宣读或已经宣读的多篇论文、著者及其单位等。例如,要检索计算机结构和操作系统方面的论文,可以通过"数学和计算机科学"(Mathematical and Computer Science)类目及其相应的著录款目,得到由ACM和IEEE主办的一次会议及其论文集中所有的文章信息。下面举一个款目例子进行说明。

892 0291①;3rd International Conference on Architectural Suppod for Programming Languages and Operating Systems(ASPLOS Ⅲ)② 3—6 Spr 1989③ Boston,MA(USA)④ Association for Computing Machinery(ACM);IEEE Computer Society⑤ 90—007804⑥ Architecture and compiler tradeoffs for a long instruction word microprocessor⑦ R. Cohn,T. Gross,M. Lam,P. S. Tseng(Dep. Comput. Sci.,Canegie—Mellon Univ.,Pittsburgh,PA).⑧ no. 1,pp. 387—397⑨

说明:①为会议登记号;②为会议名称;③为会议日期;④为会议地点;⑤为主办单位;⑥为论文顺序号;⑦为论文名称;⑧为著者名称及单位;⑨为其他补充信息,如论文

页数或参考文献数等。

CPI 的索引体系分期索引和年度累积索引两种。主要包括主题索引(subject index)、著者索引(author index)、会议日期索引(index by date of conference)、会议地点索引(index by conference location)和会议议题分类索引(index by topic of conference)等。

3.《科技会议录索引》

《科技会议录索引》(Index to Scientific&Technical Proceedings,ISTP)是一种综合性的科技会议文献检索刊物,1978 年创刊,月刊。ISTP 覆盖的学科范围广,收录会议文献齐全,出版速度快,其声誉已超过其他同类刊物而成为检索正式出版的会议文献的权威性工具。就学科范围而言,ISTP 收录的会议录涵盖了农业、环境、生物化学、生物技术、医学、工程、计算机、物理等学科;就会议类型而言,ISTP 涉及一般性会议、座谈会、研究会、讨论会、发表会等;就出版速度而言,ISTP 出版比较及时,时差仅为 6～10 周。1998 年,ISI 进一步推出基于 Web of Knowledge 平台的 ISTP 的 Web 版,极大地提升了 ISTP 的更新速度和服务水平。通过 ISTP 不仅可以快速有效地查找某个会议的主要议题和内容,而且还能够根据它所提供的会议论文作者的详细地址,直接写信向作者索取文献资料。ISTP 有月刊和年度累积本两种形式,全部内容由七个部分组成,其小类目索引是正文的编排根据,会议录目录是正文,其他则是各种索引。交叉学科的会议录在相关的学科主题下相互参见。

《科技会议录索引》即 ISTP 有四个重要作用:①取得确切的目录;②提供最新的知识;③进行回溯检索;④取得已出版的会议录。每月一期的 ISTP 都能够提供有关最近出版的会议录的信息。通过查阅每月的 ISTP。用户就能及时了解到与他的专业有关的会议文献,从而避免了不必要的重复;ISTP 是一种很容易使用的查找最新知识的工具。通过浏览其中每个会议录和每篇论文,可在很短时间里发现与自己有关的项目,当没有时间来浏览每月目次表时,则可使用 ISTP 的几个索引,准确地查出与工作者有关的那些会议录和论文。

科学技术会议录索引简称 ISTP,又称为 CPCI,被列入“三大文献索引”之一,它的网络版就是 Conference Proceedings Citation Index(CPCI),美国科学情报研究所(ISI)基于 Web of Science 的检索平台,将 ISTP(科学技术会议录索引)和 ISSHP(社会科学及人文科学会议录索引)两大会议录索引集成为 ISI Proceedings。集成之后 ISTP 分为文科和理科两种检索,分别是 CPCI-SSH 和 CPCI-S。所以它们还统称为 ISTP,也有人称它们为 CPCI。

系统提供 Full Search 和 Easy Search 两种检索界面。Full Search:提供较全面的检

索功能，通过主题词、作者名、期刊名、会议或作者单位等途径检索，可限定检索结果的语种、文献类型、排序方式，可存储/运行检索策略。Easy Search：检索功能相对简单，可以对感兴趣的特定主题、人物、地点进行检索。

1）全面检索

Full Search（全面检索）进入数据库后，单击 Full Search 按钮进入 Full Search 检索界面。检索前先进行选择。

（1）选择数据库。科学技术会议录索引（Science & Technology Proceedings）或社会科学及人文科学会议录索引（Social Sciences & Humanities Proceedings），默认为两库都选。

（2）选择年代范围。可以选择某年或最近几周上载的数据，默认为 All years。

单击 General Search 按钮进入检索词输入界面后，根据需要在以下五个字段中输入检索词，检索词间可用逻辑算符（AND、OR、NOT、SAME）连接。

TOPIC：主题词，在文献篇名、文摘及关键词字段检索，也可选择只在文献篇名（title）中检索。

AUTHOR：作者姓名，标准写法为姓氏全拼＋名的缩拼。如检索张小东就输入 zhang xd。

SOURCE TITLE：来源出版物全名。

CONFERENCE：会议信息，例如，会议名称、地点、日期、主办者，如 AMA and CHICAGO and 1994。

ADDRESS：作者单位或地址。例如，输入 IBM SAME NY 检索作者地址为 IBM's New York facilities 的会议文献。

（3）检索符几点说明如下。

① 截词符为 * ，例如输入 automat * 可以检索到 automation、automatic 等词。

② 作者单位名称常常用缩写，例如 Univ Sci & Technol Beijing，如果不能确定缩写名称，可以用 univ * and Beijing and tech * 等来检索。

③ 逻辑算符 SAME 表示检索词出现在一句话中。

（4）输入检索词后，单击 Search 按钮检索，单击 Clear 按钮清除输入框中所有内容。

（5）检索结果限定过滤。Full Search 方式还在输入框下方提供三组限定选项。

文献语种选项——默认为所有语种“All Languages”。

文献类型选项——默认为所有文献类型“All document types”。

命中结果排序选项——可以根据收录日期、相关性、第一作者姓名字顺、来源出版物

名称字顺、会议名称字顺排序。默认为“Latest Date”，即根据文献的收录日期排序。

2）简单检索

与 Full Search 类似，首先选择数据库范围，然后选择需要查找的信息类型：主题（topic）、人物（person）、地点（place），分别进入各自的检索界面。

（1）Topic Search（主题检索）：在篇名、文摘及关键词字段通过主题检索文献。步骤如下。

输入描述文献主题的检索词，用逻辑算符（AND、OR、NOT）连接。

选择结果排序方式——Relevance（相关度）或 Reverse chronological order（年代倒序）。

（2）Person Search（人物检索）：对特定人物进行检索。步骤如下。

① 输入要检索的人名，标准写法为姓氏全拼＋名的缩拼。如检索张小东就输入 zhang xd。

② 选择是检索该人物撰写的文献还是有关该人物的文献记录。

（3）Place Search（地址检索）：从著者所在机构或地理位置角度进行检索。步骤是直接输入著者所在机构（如大学或公司名称中的关键词）或地理位置（如国别或邮编），单击 Search 按钮开始检索。

12.2.4　国内会议文献的检索

1. 中国学术会议文献数据库

中国科技信息研究所主办，1982 年创刊，原名《国内学术会议文献通报》，1987 年改为现名，月刊。《中国学术会议文献通报》的报道范围广泛，几乎涵盖了自然科学、工程技术、社会科学、管理科学、农业科学和医学等所有学科，是目前报道在我国召开的国际性和全国性学术会议及会议文献的最具权威性的检索工具。年报道会议 1000 个左右，年报道量约 2 万条。《中国学术会议文献通报》是一种综合性的检索工具，由“文献通报”、“会议预报”和“会议动态”三个相互独立的部分组成。

“文献通报”是《中国学术会议文献通报》的主体，也是检索会议文献的主要工具。它分类进行编排，大类下列出包含该类内容的所有会议；会议下面再列出在该会议上交流的所有论文。其著录内容包括会议名称、会议时间、地点、会议主办单位、《中图法》分类号、会议论文编号、论文篇名、论文著者及所在单位、论文集名称、编者、出版年月、论文在论文集中的起止页码和馆藏索取号等。其中，会议论文编号由八位数字组成，前两位代表年代，后六位是本年度的流水号，论文篇名用黑体字印刷。

目前中国学术会议文献通报已经建成中国学术会议文献数据库(China Conference Paper Database,CCPD),收录始于1983年,4000个重要的学术会议,年增20万篇全文,每月更新,国家级学会、协会、部委、高校召开的全国性学术会议为主,国内目前收录会议数量较多、质量较高、学科覆盖较广。见图12-17。我国万方会议文献数据库总量达306万多篇。会议文献分为两大类:一是学术会议分类,二是主办单位分类。

(1)资源标引:采用受控语言进行主题标引,以《汉语主题词表》为叙词表,按照《中国图书资料分类法》分类。

(2)特色:收录会议级别高,全国重点会议(会议名称包含"国际"、"中国"、"多边"、"双边"、"全国"等)数量占收录会议总量的90%以上;是国内目前收录会议数量较多、学科覆盖较广的数据库 ;收集年代久远,有些机构、专业的会议已形成系列;同时收录中文与西文会议,使资源更加丰富、完整。提供的会议文献检索方法有文献题名、关键词、摘要、作者、作者单位、会议名称及其主办单位。



图12-17 万方CCPD会议文献检索与资源分类实例

2. 中国重要会议论文全文数据库

(1)概述。中国重要会议论文全文数据库是CNKI的重要服务产品之一,该数据库收录我国2000年以来国家二级以上学会、协会、高等院校、科研院所、学术机构等单位的会议论文集,年更新约10万篇论文。至2016年5月,累积中外文会议论文全文文献近210万篇。会议文献来源包括中国科协及国家二级以上学会、协会、研究会、科研院所、政府举办的重要学术会议、高校重要学术会议、在国内召开的国际会议上发表的文献。

数据库将会议文献产品分为十大专辑：基础科学、工程科技Ⅰ、工程科技Ⅱ、农业科技、医药卫生科技、哲学与人文科学、社会科学Ⅰ、社会科学Ⅱ、信息科技、经济与管理科学。十大专辑及其对应的会议文献总量如图 12-18 所示。

（2）导航检索。导航检索主要通过四级导航目录来实现递进式检索。一级导航有会议导航、论文集导航和主办单位导航；二级导航是在一级导航的基础上划分为三大子类即学科导航、行业导航和党政导航；三级导航为二级的细化，例如在“会议导航＞学科导航＞基础科学(1461)”中，就是三级导航的“基础科学”有 1461 个会议，实例如图 12-19 所示。



图 12-18　中国重要会议论文全文数据库产品分类及其文献总量

（3）检索方式丰富。主要有快速检索、标准检索和专业检索三大类，以及作者检索、基金检索、句子检索、来源会议检索等辅助检索形式。见图 12-20。



图 12-19　中国重要会议论文全文数据库的导航检索实例

① 快速检索。针对会议文献数据库的初级用户或较陌生的用户，检索时只需要用户输入简单的检索项即可，不需要做数据库选择或检索词的逻辑组配。快速检索的检索结果准确性较低。

② 标准检索。针对数据库检索应用比较熟练的用户，检索结果的准确性较高。需要对检索会议文献的会议时间、会议名称、会议级别（国际性、全国性、地区性等）、支持基金、论文集类型、语种、作者、作者单位等进行细化与过滤。

图 12-20 中国重要会议论文全文数据库的丰富检索方式

③ 专业检索。专业检索针对数据库应用的高级用户或数据库检索服务的专业人员，需要拟定科学合理的检索表达式。

可检索字段：SU＝主题，TI＝篇名，KY＝关键词，AB＝摘要，FT＝全文，AU＝作者，FI＝第一作者，AF＝作者单位，CV＝会议名称，CP＝论文集名称，RF＝参考文献，CT＝会议时间，，RT＝更新日期，FU＝基金，CLC＝中图分类号，SN＝ISSN，CN＝统一刊号，IB＝ISBN，CF＝被引频次。现在举例如下。

例一，TI＝'生态'and KY＝'生态文明' and（AU ％ '陈'＋'王'）可以检索到篇名包括"生态"并且关键词包括"生态文明"并且作者为"陈"姓和"王"姓的所有文章。

例二，SU＝'桂林'＊'旅游' and FT＝'环境保护'可以检索到主题包括"桂林"及"旅游"并且全文中包括"环境保护"的信息。

例三，SU＝（'经济发展'＋'可持续发展'）＊'转变'－'泡沫' 可检索"经济发展"或"可持续发展"有关"转变"的信息，并且可以去除与"泡沫"有关的内容。

3. 中国学术会议在线

"中国学术会议在线"（http://www.meeting.edu.cn）是经教育部批准，由教育部科技发展中心主办，面向广大科技人员的科学研究与学术交流信息服务平台。见图 12-21。

"中国学术会议在线"本着优化科研创新环境、优化创新人才培养环境的宗旨，针对当前我国学术会议资源分散、信息封闭、交流面窄的现状，通过实现学术会议资源的网络共享，为高校广大师生创造良好的学术交流环境，以利于开阔视野，拓宽学术交流渠道，促进跨学科融合，为国家培养创新型、高层次专业学术人才，创建世界一流大学做出积极贡献。



图 12-21　中国学术在线系统 Logo 与主要功能模块

“中国学术会议在线”利用现代信息技术手段，将分阶段实施学术会议网上预报及在线服务、学术会议交互式直播/多路广播和会议资料点播三大功能。为用户提供学术会议信息预报、会议分类搜索、会议在线报名、会议论文征集、会议资料发布、会议视频点播、会议同步直播等服务。

“中国学术会议在线”还将组织高校定期开办“名家大师学术系列讲座”，并利用网络及视频等条件组织高校师生与知名学者进行在线交流。提供会议资源的模糊检索、会议检索、视频检索和会议论文摘要检索四大类检索，实例如图 12-22 所示。



图 12-22　中国学术会议在线系统检索功能

12.3　学位论文检索

12.3.1　学位论文概述

学位论文是高等院校和科研院所的本科生、研究生为获得学位资格(博士学位、硕士学位和学士学位)而撰写的学术性较强的毕业研究论文，英国称为“Thesis”，美国称为“Dissertation”。学位论文通常都是经过悉心指导，符合授予学位的要求，不少论文选题新颖，论述系统，见解独到，具有独创性，特别是博士学位论文，探讨一些前人没有论及过的新领域，并且提出具有独特、创新的见解。因此，学位论文是学者、专家及博士与硕士生智慧的结晶，是了解国内外科技研究发展的重要的信息媒介，是各国拥有自主知识产权的重要信息资源和知识宝藏，具有重大的开发利用价值。

学位论文除在学位授予单位被收藏外，一般还在国家指定单位专门进行收藏。国内收藏硕士、博士学位论文的指定单位是中国科学院技术信息研究所和国家图书馆。另外，设有硕士和博士教学点的大学或研究所也藏有本校（本所）攻读硕士学位和博士学位的学位论文。按照中国高等教育文献保障体系（CALIS）要求，各高校学位论文要数字化并上网，这为学位论文的检索带来了极大方便。

12.3.2 国外重要学位论文数据库检索

国外博士硕士论文数据库 ProQuest Digital Disserationas (PQDD)是美国 UMI 公司 PorQuest Direct（PQD）系统的博硕士论文题录与文摘数据库，是 DAO（Dissertation Abstracts Oddisc）的网络版，该库收录了欧美 2 000 余所大学的 200 多万篇学位论文，ProQuest 公司是世界上最早及最大的博硕士论文收藏和供应商，该公司的 ProQuest Dissertations and Theses（PQDT）数据库收录有欧美 2 000 余所大学的 200 多万篇学位论文。国内若干图书馆、文献收藏单位每年联合购买一定数量的 ProQuest 学位论文全文，提供网络共享，即凡参加联合订购成员馆均可共享整个集团订购的全部学位论文资源，PQDT 也是世界上最大和最广泛使用的学位论文数据库，内容覆盖理工和人文社科等领域。PQDT 学位论文检索系统主界面见图 12-23。

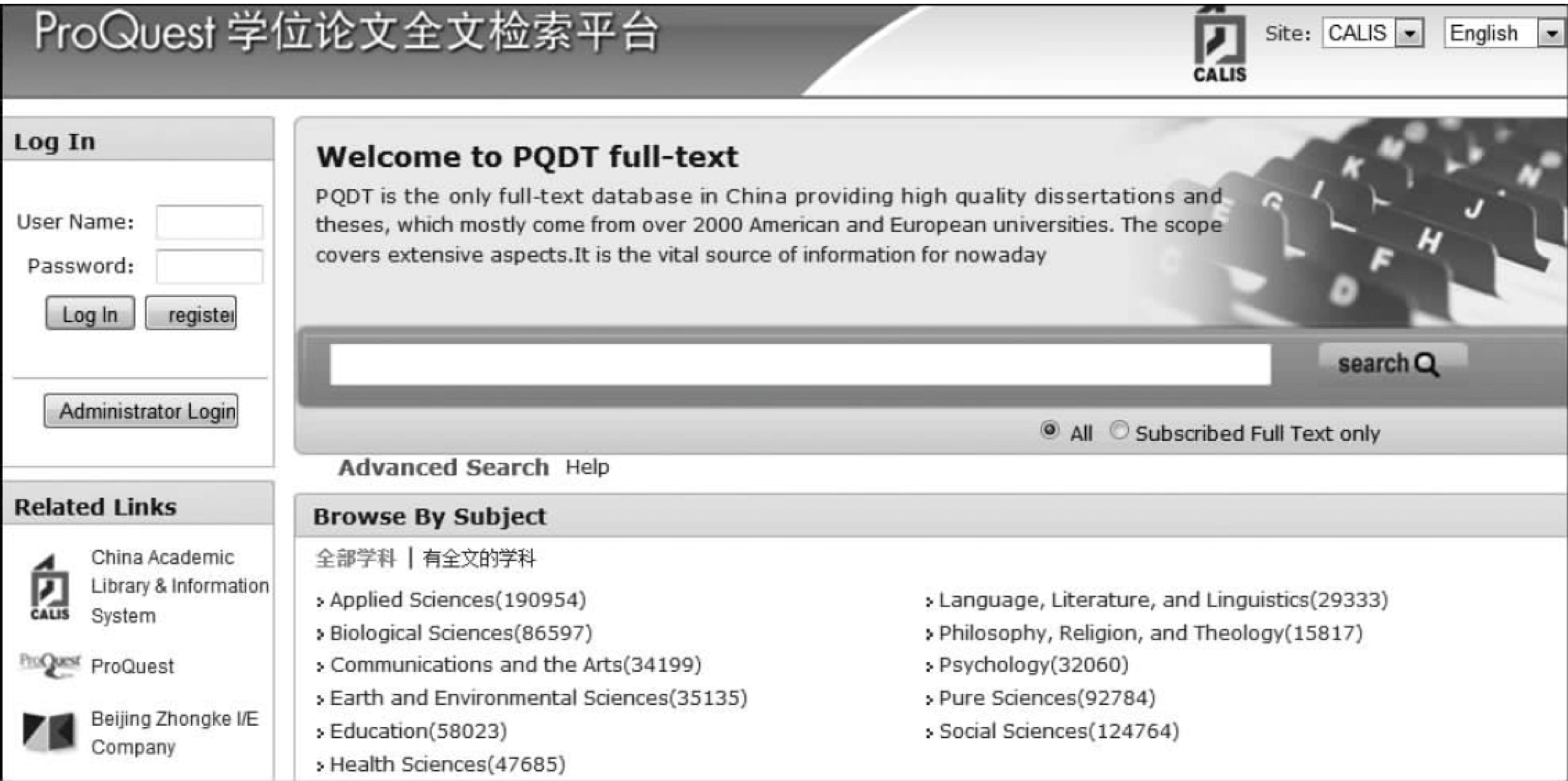


图 12-23 PQDT 学位论文检索系统主界面

1. 基本检索方法

（1）使用“and”或空格搜索全部关键词。多个关键词用空格或“and”隔开，如“digital library”或“digital and library”，这两个词将同时出现在标题、正文或摘要中（两个词的出现位置不一定相邻）。

（2）使用双引号搜索完整的关键词。如果输入的关键词本身包括空格并且不希望被分隔，可以在关键词两边加上西文双引号，如“digital library”。

（3）使用“OR”搜索任意关键词。搜索多个关键词中的任一词，如“digital or library”，这时搜索结果将包含这两个词中的任一个或全部。

（4）使用“and not”排除关键词。排除包含指定关键词的搜索结果，如“digital library and not study”，这时搜索结果将同时包含前两个词，但不包含“study”。

2. 高级检索

学位论文的高级检索包括论文标题、摘要、学科、作者、单位、导师、来源、出版时间、学位等级、语种、ISBN 之间的选择过滤及它们之间的检索逻辑表达（逻辑与、或、非）。见图 12-24。



图 12-24　PQDT 学位论文检索系统高级检索界面

12.3.3 重要国内学位论文数据库检索

1. 中国学位论文数据库

中国学位论文数据库资源由国家法定学位论文收藏机构中国科技信息研究所提供，并委托万方数据加工建库，收录了自1977年以来我国各学科领域的博士学位、硕士学位、研究生论文。《中国学位论文全文数据库》精选相关单位近些年来的博硕论文，涵盖自然科学、数理化、天文、地球、生物、医药、卫生、工业技术、航空、环境、社会科学、人文地理等各学科领域，充分展示了中国研究生教育的庞大阵容。

(1) 检索概述。本系统为“中国学位论文全文数据库”提供了多种检索途径，包括个性化检索、高级检索、字典检索、分类检索等，以便于用户迅速检索出所需要的论文资源。

“个性化检索”入口针对具体数据资源的特点，为用户提供了一个方便易用、组配灵活的检索入口，适合所有用户使用。

“高级检索”支持布尔检索、相邻检索、右截断检索、同字段检索、同句检索和位置检索等全文检索技术，具有较高的查全率和查准率。“高级检索”功能适合对检索技术有较多了解的用户使用。

本系统灵活易用、高效强大的检索功能基于其灵活、先进的数据库索引技术。用户在利用此系统检索时，系统先利用事先建好的索引表找出符合条件的记录，再从数据库中读取相关记录。本系统通过“字段编号”识别字段，通过“索引编号”识别索引项。因此，用户限定在某一字段检索时，实际上是限定在对应索引号中检索。

此处介绍可检索字段时，会在各个字段名称后的括号内列出其对应索引项“编号”。使用本系统“高级检索”功能的用户可能需要了解这方面内容。

“学位论文全文库”的可检索字段有论文题名(200)、作者(300)、作者专业(720)、导师姓名(380)、授予学位(700)、授予单位(303)、授予时间(440)、分类号(610)、关键词(620)、文摘(600)。

“学位论文全文库”中支持“精确匹配”检索的字段有作者(300)、作者专业(720)、导师姓名(380)、授予学位(700)、授予单位(303)、分类号(610)、关键词(620)。

(2) 个性化检索。“个性化检索”入口针对具体数据资源的特点，为用户提供了一个方便易用、组配灵活的检索入口，适合所有用户使用。在利用“个性化检索”入口检索时，用户只需通过下拉菜单单击所要检索的字段，输入相应检索词，便可组配出比较复杂的检索表达式，查找出相关信息。“中国学位论文全文数据库”(以下简称学位论文库)的个性化检索入口如图12-25所示。

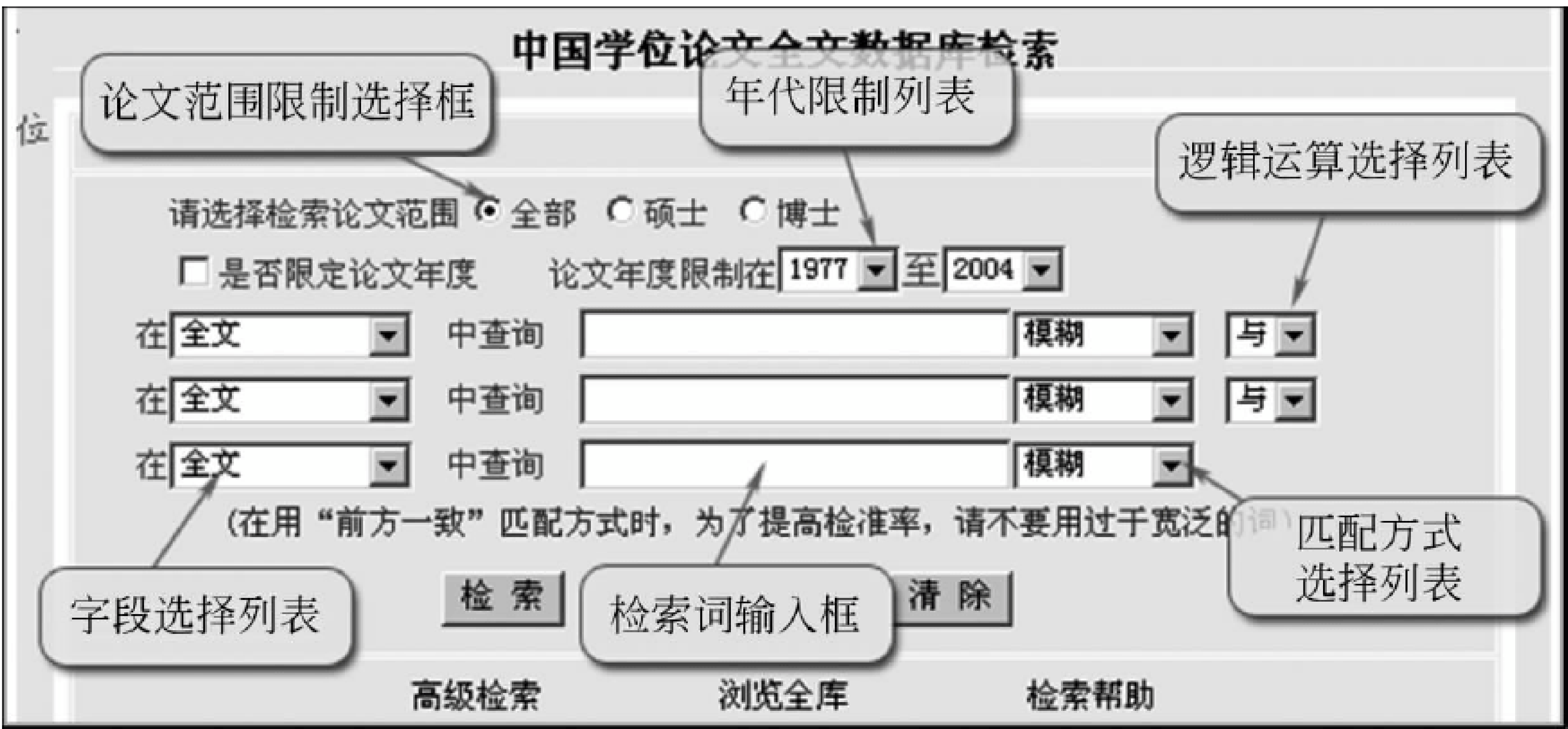


图 12-25　中国学位论文文摘数据库个性化检索界面

(3) 二次检索。二次检索是在已有检索结果范围内再一次检索，以便进一步缩小检索范围。

“学位论文库”的检索结果显示格式如图 12-26 所示。此页面的上方提供了二次检索入口，其使用方法与“个性化检索”入口相同。在此，用鼠标单击此页面下方的“显示选择记录”按钮，便可按“选择显示格式”栏所指定的显示格式浏览选定记录（记录前的方框中有钩的记录为选定记录）。



图 12-26　中国学位论文文摘数据库二次检索实例

(4) 关联检索。在一记录的全部信息中，不仅会以适当形式给出此记录的相关信息，还有可能提供关联检索入口。对全文数据库，还会提供访问对应的全文链接。如图 12-27 所示，学位论文库的“全部信息”显示格式中不仅给出了论文的相关信息，还提供了查看论

文全文的链接与“关联检索”入口。

WebParallel:一种新型的并行计算模型的设计与实现

单击链接，浏览该专业所有论文

【论文题名】WebParall 行计算模型的设计与实现
【论文作者】高峰
【专业名称】计算机软件与理论
【导师姓名】徐甲同
【授予学位】硕士
【授予单位】西安电子科技大学
【出版时间】20000101
【 分类号 】TP312 TP38
【 关键词 】Java小应用程序 浏览器 运行支持系统 并行计算 WebParallel
【 馆藏号 】Y340993
【论文页数】54页
【文摘语种】中文文摘
【 文 摘 】该文是研究在Web计算环境下 问题.研究人员提出了一种新的计算模型:WebParallel,它以浏览器作为并行节点机,Java小应用程序作为并行任务单元,用户根据自愿的原则加入并行计算.WebParallel模型分为两层:上层为虚拟机层,为程序员提供了一个同构的编程环境,减轻了程序员设计并行应用程序的负担;下层为运行支持系统(runtime system)层,负责把虚拟机映射到实际的Web环境上.在论文中,针对并行计算和Web的特点,研究人员重点研究了四个问题:1.共享数据的存储问题;2.并行任务的通讯问题;3.并行应用的分布问题;4.并行计算的负载调度问题.研究人员实现了该模型并在模拟环境下进行了实验,证明该模型有较好的应用价值.

单击链接，浏览属于该分类的所有论文

单击链接，浏览含有该关键词的所有论文

查看全文

单击链接，浏览论文全文

图 12-27 中国学位论文文摘数据库关联检索实例

专业名称：单击专业名称，可检索出此“学位论文库”中“专业名称”为此专业的所有论文。

导师姓名：单击导师姓名，可检索出此“学位论文库”中“导师姓名”为此姓名的所有论文。

授予单位：单击授予单位，可检索出此“学位论文库”中“授予单位”为此单位的所有论文。

分类号：单击一分类号，可检索出此“学位论文库”中此分类下的所有论文。

关键词：单击一关键词，可检索出此“会议论文库”中“关键词”中有这个词的所有论文。

查看全文：单击“查看全文”这个链接，可以查看这篇论文的全文。

2. CNKI 中国优秀博硕士学位论文数据库

CNKI 中国优秀博硕士学位论文数据库是目前国内相关资源最完备、高质量、连续动

态更新的中国优秀博硕士学位论文全文数据库。目前,累积博硕士学位论文全文文献近300万篇(从1984年至今的博硕士学位论文)。覆盖基础科学、工程技术、农业、医学、哲学、人文、社会科学等各个领域,收录全国426家培养单位的博士学位论文和699家硕士培养单位的优秀硕士学位论文。产品分为十大专辑:基础科学、工程科技Ⅰ、工程科技Ⅱ、农业科技、医药卫生科技、哲学与人文科学、社会科学Ⅰ、社会科学Ⅱ、信息科技、经济与管理科学。十大专辑下分为168个专题。

(1) CAJ Vierwer 专门浏览器与全文浏览。CAJ 为中国学术期刊全文数据库的英文缩写(China Academic Journals),CAJ Vierwer 是 CNKI(中国知识基础设施工程)资源的专门全文浏览器。其标识见图 12-28。



图 12-28 CAJ Vierwer 浏览器标识

CAJ Vierwer 阅读器是光盘国家工程研究中心、清华同方知网(北京)技术有限公司的系列产品,它支持中国期刊网的 CAJ、NH、KDH 和 PDF 格式文件。它可以在线阅读中国期刊网的原文,也可以阅读下载到本地硬盘的中国期刊网全文。主要全文阅读功能有以下几项:

① 页面设置:改变文章原版显示的效果,可以设置两种页面显示方式,即对开显示及连续对开显示。

② 浏览页面:实现页面的任意跳转,页面内容旋转与标注。

③ 查找文字:对于非扫描文章,提供全文字符串查询功能。

④ 切换显示语言:除了提供简体中文外,还提供了繁体中文、英文显示方式,方便海外用户使用。

⑤ 文本图像摘录:实现文本及图像摘录并可将摘录结果粘贴到 WPS、WORD 等编辑器中进行任意编辑。

⑥ 打印及保存：将可查询到的文章以 * . caj、. kdh、. nh、. pdf 文件格式保存，并可将其按照原版显示效果打印，可以打印预览或设置书面打印。

⑦ 内容转换：可以将. caj 格式的内容转换为. word 或. wps 格式的内容。

CAJ Vierwer 浏览器的内容目录查阅实例见图 12-29。

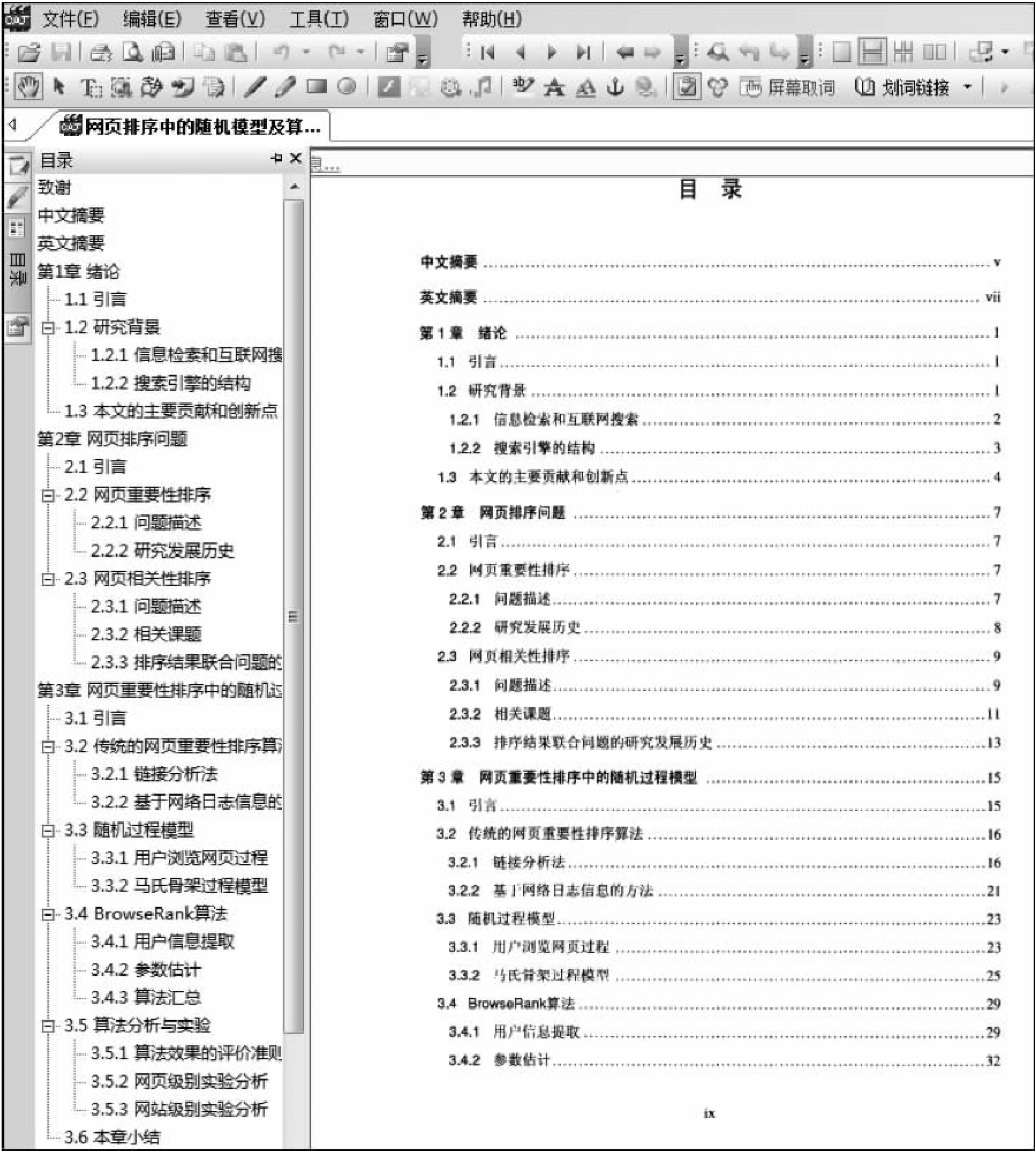


图 12-29 CAJ Vierwer 浏览器的内容目录查阅实例图

(2) 检索类型。CNKI 中国优秀博硕士学位论文数据库与其他的 CNKI 资源库(例如期刊库等)一样提供有基本检索、高级检索、专业检索、基金检索和句子检索五种检索。见图 12-30。



图 12-30　中国优秀博硕士学位论文数据库的检索类型示意图

检索条件即检索项或检索式的基本信息查询与过滤条件包括：

① 基本检索项包括学位论文的主题、题名、作者、导师、授予单位、关键词、摘要、目录、全文、参考文献、分类号、学科专业等。

② 检索词之间的布尔逻辑关系：并含、或含与不含。

③ 时间范围限定，在 1980 年到 2016 年之间任意选择一个时间点或时间段。

④ 论文级别的查询过滤：国家级优秀论文、省级优秀论文和校级优秀论文。

其他的高级检索（针对熟练用户）、专业检索（针对专门信息服务人员、信息分析人员）、基金检索（针对某一基金项目的研究内容聚类）和句子检索（查询包含两个关键词的句子，实现对事实的检索）的方法大致相近。

(3)高级检索。主要针对复杂检索主题，便于获得准确度高（即高查准率）的检索结果。高级检索可以对高达八个主要检索项（主题、题名、关键词、摘要、全文、参考文献、分类号和学科专业）进行逻辑组合检索以及同时对七个辅助检索项进行逻辑查询（学位授予时间、数据更新时间、授予单位、作者、作者单位、支持基金和优秀论文等级）。见图 12-31。

(4) 检索结果的排序。对于检索结果，可以依据查询的主题相关度、发表时间顺序、引用量、下载量、学位授予时间来排序。例如以“网页排序”或者“PageRank”为主题词（2 个词的中英文意义相同，所以逻辑式为或）且时间范围过滤为“2005—2016 年”，检索到需求论文 62 篇。依据论文的下载量对结果进行排序，部分检索实例如图 12-32 所示。

(5) 检索结果内容的全文下载与阅读。对于选择的检索结果内容既可以在线全文阅读，也可以对学位论文分页、分章或整本下载。下载后的论文内容需要在其专用阅读器环境下阅读使用。见图 12-33。



图 12-31　中国优秀博硕士学位论文数据库的高级检索界面



图 12-32　中国优秀博硕士学位论文数据库的检索结果排序部分实例图



图 12-33　中国优秀博硕士学位论文数据库的检索结果内容下载与阅读模块提示

（6）学位论文文献间的引用网络图谱。学位论文文献间的引用网络图谱（见图 12-34）揭示了某一研究主题相关的学位论文文献之间的研究价值脉络与相互关联性影响。这对于大学生而言，特别是研究生进行探究性学习和研究性学习有重要参考价值。这不仅使得大学生用户能够利用它逐步把握某一研究主题的研究进程与研究者们的相互影响关系，也是创新思维启发或进一步深入研究的重要基础。

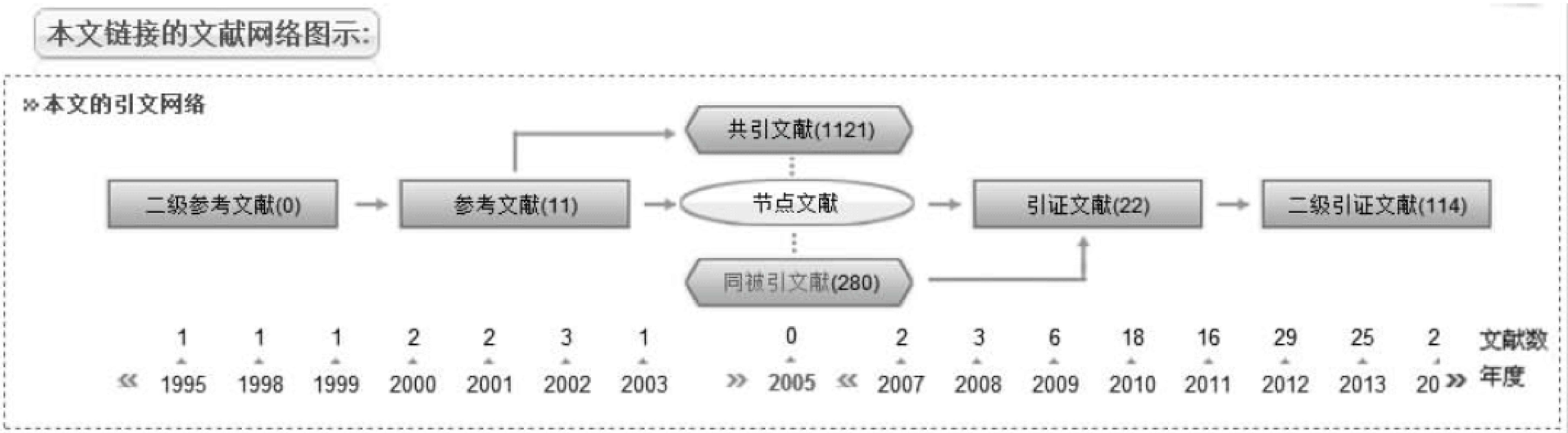


图 12-34　中国优秀博硕士学位论文数据库的文献间网络关系

① 参考文献：指在学术研究过程中对某些文献的整体性参考与借鉴。

② 共引文献：共引文献也称同引文献，是指与本文有相同参考文献的文献，与本文有共同研究内容。共引文献数量越多，文献间的相关性越大。

③ 同被引文献：是指与本文同时被作为参考文献引用的文献，与本文共同作为进一步研究的基础。如果 A、B 两篇文献均被 C 文献作为参考文献引用，则文献 A 与文献 B 存在同被引关系。

④ 引证文献：是指引用本文的文献，是本文研究工作的继续、应用、发展或评价。引证文献是学术论著撰写中不可或缺的组成部分，也是衡量学术著述影响大小的重要因素。作者的文献被引证的次数越多，此作者的文献越有价值。

⑤ 二级参考文献：本文参考文献的参考文献。

⑥ 二级引证文献：本文引证文献的引证文献。

12.4 专利文献资源检索

专利文献是科学技术的宝库。它融技术、法律和经济信息于一体，是各单位各部门领导了解掌握国内外技术发展现状，进行技术预测和做出科学决策的依据，是科研人员和工程技术人员进行课题研究，解决技术难题不可缺少的工具；是发明人寻找技术资料，不断做出新的发明创造的源泉。在技术贸易中，专利文献可用于了解专利技术的法律状态；在技术和市场竞争中，专利文献可用于判定侵权行为；在申报国家发明成果奖和申请专利时，专利文献可用于确定其新颖性、创造性。企业可利用专利文献了解和监视同领域竞争对手的情况，开发适销对路的新产品。专利文献可以为国家经济建设服务，为各单位增加竞争与发展活力服务。

12.4.1 专利与专利文献概念

专利（patent）一词包含三层含义：一指专利法保护的发明创造与设计，二指专利权，三指专利说明书等专利文献。其核心是一种法律制度，即专利制度，而专利权和专利文献是专利的具体体现。

专利权是知识产权的一种。作为一种无形财产，专利权具有专有性、地域性和时间性。知识产权是人们利用知识获得成果的专有权，是相对实物产权而言的，所以也叫智力成果权。知识产权受法律保护，任何人未经知识产权所有人的许可，不准使用、制造或销售其成果，否则就构成侵权行为，并受到法律的制裁。知识产权包括工业产权和版权两部分。工业产权是涉及工业、农业、商业、采掘业和一切制造成品或天然产品的产权，包括专利、商标、服务标记、厂商名称、货源名称或原产地名称和制止不正当竞争等。版权也称著作权，指作者或出版者对其作品享有印刷、出版、复制和销售等权利。

专利文献（patent literature）是指记录有关发明创造信息的文献。广义包括专利申请书、专利说明书、专利公报、专利检索工具以及与专利有关的一切资料；狭义仅指各个国家或地区的专利局出版的专利说明书或发明说明书。

12.4.2 专利文献的类型与作用

1. 专利文献的主要类型

（1）按专利的实质内容划分。由于世界各国的专利法不同，专利种类的划分也不尽相同。美国分为发明专利、外观设计专利和植物专利。中国、日本、德国等国分为发明专

利、实用新型专利和外观设计专利。发明专利是国际上公认的应具备新颖性、先进性和实用性的新产品或新方法的发明;实用新型专利是对机器、设备、装置、器具等产品的形状构造或其结合所提出的实用技术方案;外观设计专利是指对产品的外形、图案、色彩或其结合做出的富有美感而又适于工业应用的新设计。实用新型专利和外观设计专利都涉及产品的形状,两者的区别是:实用新型专利主要涉及产品的功能,外观设计专利只涉及产品的外表。如果一件产品的新形状与功能和外表均有关系,申请人可以申请其中一个,也可分别申请。

(2) 按专利刊载的形式划分:专利申请书、专利说明书、专利公报、专利检索工具、专利分类表、与专利有关的法律文件及诉讼资料等。其中尤为重要的是专利说明书和专利公报。

专利说明书是专利文献的主体。它是个人或企业为了获得某项发明的专利权,在申请专利时必须向专利局呈交的有关该发明的详细技术说明,包括经审查批准的审定说明书、经审查但尚未批准的展出说明书和未经审查的公开说明书(专利申请书)。专利说明书的作用是公开新发明创造的技术内容,限定专利权保护的范围。因此,专利说明书的内容主要涉及的就是发明创造的技术内容和权利内容。

各国的专利说明书都有固定的格式,一般由三部分组成:一是著录项目(标头),包括专利号、专利申请号、申请日期、公布日期、专利分类号、发明题目、专利摘要、专利权范围、法律上有关联的文件、专利申请人、专利发明人、专利权所有者等。每个著录事项前通常有国际通用的数据识别代号(INID)。二是发明说明书(正文),是申请人对发明技术背景、发明内容以及发明实施方式的说明,常常附有插图。三是专利权项(权项或权利要求书),是专利申请人要求专利局对其发明给予法律保护的项目,当专利批准后,权项具有直接的法律作用。

2. 专利文献的主要作用

(1) 对专利申请进行专利性检索。申请人在申请专利前,应检索相关的专利文献,看看该项发明是否具有新颖性、创造性与实用性,以免提出申请后不能获得专利权;发明专利的申请人请求实质审查,按专利法规定应向专利局提交相关的参考资料,包括专利文献。

(2) 启迪发明创造思路。"站在巨人的肩膀上"就是专利利用的重要名言,许多发明是从他人的发明基础上发展起来的,或者从中获得启发、借鉴。

(3) 可以了解某领域的最新动态。专利文献的报导比其他文献早 1～3 年,而且一项新技术的诞生到推广应用有个过程,存在一个时间差,少则几个月,多则几十年。因此我

们从专利文献中可以了解科技发展的最新动态。

(4) 有利于技术转让。企业科技工作者在寻找新技术时，无非是两种途径：一是企业主动出击，二是发明人毛遂自荐。对于主动出击，较好的方法是检索专利文献，在该技术领域中检索出众多的技术，然后择优筛选；对于毛遂自荐，更应检索专利文献，可以避免被自荐者的发明特点所迷惑。

(5) 有利于企业的技术开发。从以往的教训来看，许多企业盲目研制一些新产品，不仅造成人力、物力、财力的浪费，而且可能与以往的技术相比，并不是先进的技术，结果其产品在市场上销售不畅。进行专利检索，可以避免浪费和重复劳动，而且可以借鉴以往的发明，开发出技术先进且有市场潜力的产品；同时还可以从中了解竞争对手的发展动态，以便采取相应的应对措施。

(6) 有利于引进国外先进技术和设备。从以往的引进来看，存在不少弊端：盲目引进，不是引进最先进的技术，技术转让中的一些专利是过期专利，结果支付了过高的技术使用费等。通过检索专利文献，不仅可以避免上述弊端，而且可以货比三家，从中找出先进且又适合国情的技术。

(7) 作为专利诉讼的有力依据。在专利侵权诉讼中，被告在被起诉侵权时，应检索专利文献，查看一下原告的专利资料及相关的背景技术，以避免败诉；专利申请人对于专利局复审委员会做出某决定(驳回或撤销或无效或维持等)不服向人民法院起诉时，同样应检索专利文献，并提供相关的佐证资料。

12.4.3 国际专利分类

专利文献检索主要有三种途径，即分类检索途径、专利权人检索途径和序号检索途径，其中最常用的是分类检索途径。而分类检索最典型的检索工具是国际专利分类表。

《国际专利分类表》(IPC 分类)是根据 1971 年签订的《国际专利分类斯特拉斯堡协定》编制的，是目前唯一国际通用的专利文献分类和检索工具，为世界各国所必备。问世的 40 多年里，IPC 对于海量专利文献的组织、管理和检索，做出了不可磨灭的贡献。由于新技术的不断涌现，专利文献每年增长约 150 万件，目前约有 5000 万件。

另外，IPC 的建立是基于纸件专利文献的管理与检索，在计算机、通信网络等新技术快速发展的今天，它显现出一些不适应。为了让 IPC 名副其实地成为世界各国专利局以及其他使用者在确定专利申请的新颖性、创造性时进行专利文献检索的一种有效检索工具，IPC 联盟大会成员国、世界知识产权组织(WIPO)在 1999—2005 年对国际专利分类表进行了改革，将第 8 版 IPC 分成基本版和高级版两级结构。第 8 版 IPC 基本版约 20 000 条，

包括部、大类、小类、大组和在某些技术领域的少量多点组的小组。第 8 版 IPC 高级版约 70 000 条，包括基本版以及对基本版进一步细分的条目。高级版供属于 PCT 最低文献量的工业产权局和大的工业产权局使用，用来对大量专利文献进行分类。

(1) IPC 分类表共分以下九个分册。

第一分册——人类生活需要。

第二分册——作业、运输。

第三分册——化学、冶金。

第四分册——纺织、造纸。

第五分册——固定建筑物。

第六分册——机械工程、照明、加热、武器、爆破。

第七分册——物理。

第八分册——电学。

第九分册——使用指南。

(2) IPC 八大类，即：

A——人类生活需要。

B——作业、运输。

C——化学、冶金。

D——纺织、造纸。

E——固定建筑物。

F——机械工程、照明、加热、武器、爆破。

G——物理。

H——电学。

为了便于查找 IPC 分类号，每一版的 IPC，国际知识产权组织都会配套编 IPC 正式索引(officail index to the IPC)，也就是 IPC 关键司及类号对照索引。它是为了帮助用户从主题词入手，确定发明的 IPC 类号而设置的辅助性检索工具。该索引以关键词作为标目，其后给出该关键词所属技术领域的 IPC 类号。有些关键词下又进一步划分出下属关键词，用来限定说明标目的含义。IPC 只用于发明专利和实用新型专利的分类与检索。外观设计专利的分类与检索须使用《国际外观设计专利分类表》(*International Industrial Design Classification*)。

12.4.4 专利搜索引擎

专利搜索引擎是针对专利信息的特殊性而建立的专门搜索引擎。它采用先进的数据挖掘及自然语言处理技术，内置强大语义分析引擎，实现专利信息的智能化检索；采用跨库联合检索技术及中英文跨语言检索技术，实现中外专利数据库的联合统一检索及中英文混合检索，突破了广大用户因语言障碍而造成查全率、查准率的问题；提供搜索引擎式检索、表格检索、表达式检索、逻辑检索等多种检索方式，满足了不同层次用户对于专利信息检索的需求。

SooPAT（http://www.soopat.com）就是大学生所熟悉的专利搜索引擎。SooPAT立足专利领域，致力于专利信息数据的深度挖掘，致力于专利信息获得的便捷化，努力创造最强大、最专业的专利搜索引擎，为用户实现前所未有的专利搜索体验。SooPAT 拥有中国最有创造力的专利专家、信息检索专家和系统架构专家，以及众多持同一理想的志愿者和广泛支持者。SooPAT 的目标是让专利搜索平民化，让不是专利检索专家的你也能在瞬间找到所需要的专利。

1. 专利引擎检索的一般方法

SooPAT 查询简洁方便，仅需输入查询内容并回车（Enter），或单击“搜索”按钮即可得到相关资料。SooPAT 尽量让最相关的专利文献出现在最前面，方便用户更容易找到最重要、最相关的内容。

（1）搜索窍门一。多个关键词之间用空格隔开，可获得更多搜索结果。如“飞机 轮胎”比“飞机轮胎”搜索结果要多。

（2）搜索窍门二。通过申请（专利）号、公开（公告）号查询时，直接输入号码，前面不用加 ZL 或 CN。

（3）不用忽略词。SooPAT 会忽略“的”、“地”、“得”等字词，这类字词不仅无助于缩小查询范围，而且会大大降低搜索速度。这些词和字符称为忽略词。

（4）检索分词应用。在一些情况下，SooPAT 会对查询词进行适当拆分，以防止漏检，比如输入“航空航天动力”，会自动转换“航空 AND 航天 AND 动力”来搜索。如不需要 SooPAT 进行这种自动拆分，只需在查询词上加英文单引号''，比如输入'航空航天动力'，就不会再拆分开了。

（5）检索文字繁简体切换。SooPAT 运用汉字繁简自动转换系统，无论输入繁体或简体字皆可查询专利。并且可通过每页右上角的繁简体切换按钮进行整页的繁简体切换。

2. 专利引擎检索的高级方法

SooPAT 的搜索框支持各字段间组成复杂的逻辑检索式进行精确搜索，如果用户是专利信息检索行业专门人员或是熟练用户想更精确地查询专利，需要掌握以下内容。

(1)字段限定。如果需要将查询词限定在某一字段内，可在这个查询词前加上以下的字段限定符，注意，字段后用英文冒号，见表 12-1。

表 12-1　字段限定检索符号

字段限定符	字段名称	字段限定符	字段名称	字段限定符	字段名称
SQH	申请号	SQRQ	申请日期	MC	专利名称
ZY	摘要	SQR	申请人	DZ	地址
FMR	发明人	FLH	分类号	ZFLH	主分类号
GKH	公开号	GKRQ	公开日期	ZLDLJG	专利代理机构
DLR	代理人	LeiXing	专利类型		

例如，“ZY:苹果”表示查询摘要里包括“苹果”这个词的专利。

“MC: 塑料 AND FLH:C08F * ”表示查询专利名称包含“塑料”，且分类号为“C08F”的专利。

“ MC: 塑料 AND FMR:许”表示查询专利名称包含“塑料”，且发明人包含“许”的专利。

(2) 缺省符“ * ”。申请号、公开日期、公开号、分类号、主分类号、申请日期这六个字段中查询时，可使用缺省符“ * ”进行模糊搜索。

例如，“GKRQ:(* 200601)”表示查询公开日期在“2006 年 1 月”的所有专利。FLH:(* A61B)表示查询“A61B”分类号小类下的所有专利。

(3) 时间范围查询。申请日、公开日可支持时间范围查询 [开始值 TO 结束值]。

例如，“SQRQ:[2005 TO 2006]”表示查询申请日期在 2005 年与 2006 年之间的所有专利。

(4) 复杂逻辑运算。SooPAT 支持 AND、OR、NOT 以及()的逻辑运算，以空格间隔默认为 AND 关系。

例如，“MC:塑料 AND FMR:许”表示查询专利名称包含“塑料”，且发明人包含“许”的专利。

“MC:塑料 FMR:许”表示与上述查询结果一致。

“MC:塑料 杯子”表示查询专利名称包含“塑料”,且所有查询域中包含“杯子”的专利。

“MC:塑料 AND (FMR:许 OR FMR:刘)”表示搜索名称中包含“塑料”,且发明人中包含“许”或“刘”的专利。

3. 专利的表格式检索

就是以丰富的专利内容项目的表格形式提供用户检索界面,用户只需要使用部分或全部表格项就可以实现理想的专利检索。

在检索时,申请(专利)号、公开(公告)号前不用加“ZL”或“CN”。字段内各检索词之间可进行 AND、OR、NOT 运算,使用时 AND、OR、NOT 必须大写。字段内各检索词之间如以空格间隔,默认为 AND 关系。表格式检索的项目内容说明如下。

(1) 申请(专利)号。搜索时需输入完整申请号。申请号前不用加“ZL”或“CN”。字段内各检索词之间可进行 AND、OR、NOT 运算。字段内各检索词之间如以空格间隔,默认为 AND 关系。检索示例如下。

① 输入完整申请号,如已知申请号为“99111770.0”,可输入:99111770.0。

② 已知申请号为“200510011420.0”,可输入:200510011420.0。

③ 已知申请号前几位为“20051001142”,可输入:20051001142。

(2) 申请日。由年、月、日三部分组成。直接输入其年、月、日所构成的连续 8 位数字,年、月、日各数字之间不用符号间隔。字段内各检索词之间可进行 AND、OR、NOT 运算。字段内各检索词之间如以空格间隔,默认为 AND 关系。检索示例如下。

① 申请日为 2015 年 07 月 21 日,可输入:20150721。

② 申请日为 2016 年,可输入:2016。

③ 申请日为 2016 年 07 月,可输入:201607。

④ 申请日为 2013 年 08 月到 2016 年 6 月,可输入:[201308 TO 201606]。

专利搜索引擎的表格式检索界面实例见图 12-35。

(3) 名称。可输入所知的完整专利名称,也可选用合适的关键字进行模糊搜索。应尽量选用合适的关键字,以免检索出过多无关文献。字段内各检索词之间可进行 AND、OR、NOT 运算。字段内各检索词之间如以空格间隔,默认为 AND 关系。检索示例如下。

① 已知名称中包含“计算机”,可输入:计算机。

② 已知名称中包含“计算机”和“应用”,可输入:计算机 AND 应用。

③ 已知名称中包含“计算机”或“控制”,可输入:计算机 OR 控制。

☑发明 ☑实用新型 ☑外观设计 □发明授权

申请(专利)号:		例: 200510011420.0	申请日:		例: 20030122
名称:		例: 发动机	公开(公告)日:		例: 20070808
摘要:		例: 计算机 控制	公开(公告)号:		例: 1664816
分类号:		例: G06F17/30	主分类号:		例: H04B7/185
申请(专利权)人:		例: 微软公司	发明(设计)人:		例: 刘一宁
地址:		例: 北京市上地	国省代码:		例: 北京
代理人:		例: 吴观乐	专利代理机构:		例: 柳沈
权利要求书:		例: 醋	说明书:		例: 醋

SooPAT 搜索　SooPAT 分析

AND　OR　NOT　(　)

图 12-35　专利搜索引擎的表格式检索界面实例

④ 已知名称中包含“计算机”,但不包含“电子”时,可输入:计算机 NOT 电子。

(4) 公开日。由年、月、日三部分组成。直接输入其年、月、日所构成的连续 8 位数字。年、月、日各数字之间不用符号间隔。字段内各检索词之间可进行 AND、OR、NOT、TO 运算。字段内各检索词之间如以空格间隔,默认为 AND 关系。检索示例如下。

① 公开日为 2016 年 08 月 08 日,可输入:20160808。

② 公开日为 2015 年,可输入:2015。

③ 公开日为 2016 年 08 月,可输入:201608。

④ 公开日为 2013 年 08 月到 2015 年 6 月,可输入:[201308 TO 201506]。

(5) 摘要。应尽量选用合适的关键字,以免检索出过多无关文献。字段内各检索词之间可进行 AND、OR、NOT 运算。字段内各检索词之间如以空格间隔,默认为 AND 关系。检索示例如下。

① 已知摘要中包含“计算机”,可输入:计算机。

② 已知摘要中包含“计算机”和“应用”,可输入:计算机 应用。

③ 已知摘要中包含“计算机”或“控制”,可输入:计算机 OR 控制。

④ 已知摘要中包含“计算机”,但不包含“电子”时,可输入:计算机 NOT 电子。

(6) 公开(公告)号。直接输入完整的公开(公告)号。公开(公告)号前不用加“ZL”或“CN”。字段内各检索词之间可进行 AND、OR、NOT 运算。字段内各检索词之间如以空格间隔,默认为 AND 关系。检索示例如下。

① 已知公开(公告)号为“1387751”,可输入:1387751。

② 已知公开(公告)号前面几位为“13877”,可输入:13877。

(7) 分类号、主分类号。分类号可由《国际专利分类表》查得。其号码格式包括部、大类、小类、大组、小组。字段内各检索词之间可进行 AND、OR、NOT 运算。字段内各检索词之间如以空格间隔,默认为 AND 关系。检索示例如下。

① 已知分类号为“G06F17/30”,可输入:G06F17/30。

② 已知分类号起首部分为“G06F”,可输入:G06F。

③ 若检索分类号为“G06F17/30”或“G06F15/17”,可输入:G06F17/30 OR G06F15/17。

④ 如为外观设计专利,其分类号格式为两位数字-两位数字,如“06-09”,可输入:06-09。

(8) 名称。可输入所知的完整专利名称,也可选用合适的关键字进行模糊搜索。应尽量选用合适的关键字,以免检索出过多无关文献。字段内各检索词之间可进行 AND、OR、NOT 运算。字段内各检索词之间如以空格间隔,默认为 AND 关系。检索示例如下。

① 已知名称中包含“计算机”,可输入:计算机。

② 已知名称中包含“计算机”和“应用”,可输入:计算机 应用。

③ 已知名称中包含“计算机”或“控制”,可输入:计算机 OR 控制。

④ 已知名称中包含“计算机”,但不包含“电子”时,可输入:计算机 NOT 电子。

(9) 申请(专利权)人。申请(专利权)人可为个人或团体。搜索时可以写出完整的申请人名,也可以只写出一部分进行关键字模糊搜索。字段内各检索词之间可进行 AND、OR、NOT 运算。字段内各检索词之间如以空格间隔,默认为 AND 关系。检索示例如下。

① 已知申请(专利权)人为“王强”,可输入:王强。

② 已知申请(专利权)人为“微软公司”,可输入:微软公司。

③ 已知申请(专利权)人名字中包含“宁”,可输入:宁。

④ 已知申请(专利权)人名字中包含“刘”和“宁”,可输入:刘 宁。

⑤ 已知申请(专利权)人为北京某塑料厂,可输入:北京 塑料。

⑥ 已知申请(专利权)人中包含“微软公司”或“西门子”,可输入:微软公司 OR 西门子。

(10) 发明(设计)人。可以写出完整的发明(设计)人名,也可以只写出一部分进行关键字模糊搜索。字段内各检索词之间可进行 AND、OR、NOT 运算。字段内各检索词之间如以空格间隔,默认为 AND 关系。检索示例如下。

① 已知发明(设计)人为“袁隆平”,可输入:袁隆平。

② 已知发明(设计)人名字中包含“宁”,可输入:宁。

③ 已知发明(设计)人名字中包含“刘”和“宁”,可输入:刘 宁。

④ 已知发明(设计)人中包含袁隆平和邓启云,可输入:袁隆平 邓启云。

⑤ 已知发明(设计)人中包含袁隆平或邓启云,可输入:袁隆平 OR 邓启云。

(11) 地址。支持模糊检索,模糊检索时应尽量选用合适关键字,以免检索出过多无关文献。字段内各检索词之间可进行 AND、OR、NOT 运算。字段内各检索词之间如以空格间隔,默认为 AND 关系。检索示例如下。

① 已知地址中包含北京市,可输入:北京市。

② 已知地址中包含北京市和中关村,可输入:北京市 中关村。

③ 已知地址中包含北京市或苏州市,可输入:北京市 OR 苏州市。

(12) 专利代理机构。可以写出完整的专利代理机构名称,也可以只写出一部分进行模糊搜索。字段内各检索词之间可进行 AND、OR、NOT 运算。字段内各检索词之间如以空格间隔,默认为 AND 关系。检索示例如下。

① 已知专利代理机构名称中包含“柳沈”,可输入:柳沈。

② 已知专利代理机构名称中包含“贸易”和“专利”,可输入:贸易 专利。

③ 已知专利代理机构名称中包含“柳沈”或“贸易促进委员会”,可用:柳沈 OR 贸易促进委员会。

(13) 代理人。可以写出完整的代理人名,也可以只写出一部分进行模糊搜索。字段内各检索词之间可进行 AND、OR、NOT 运算。字段内各检索词之间如以空格间隔,默认为 AND 关系。检索示例如下。

① 已知代理人为“吴观乐”,可输入:吴观乐。

② 已知代理人名字中包含“吴”和“乐”,可输入:吴 乐。

③ 已知代理人中包含“吴观乐”或“许鸣石”,可输入:吴观乐 OR 许鸣石。

4. 专利搜索引擎的分类搜索

分类检索包括:输入关键词查分类号和输入分类号查含义,同时 IPC(国际专利分类号)和 IDC(国际外观专利分类号)可以自由切换。也可直接在分类类目中选择需要的专利信息,例如“控制、信号”。如果进入“控制、信号”类目中,将同步展示 IPC 分类号与分类类目,以更加清晰的多级类目形式(一级、二级、三级等)展示更加丰富的专利分类导航,也可以在“中国专利”和“世界专利”之间切换查询,实例见图 12-36 和图 12-37。



图 12-36　专利搜索引擎的分类搜索主界面



图 12-37　专利搜索引擎的 IPC 分类查询实例

5. 专利搜索引擎的普通检索

普通检索针对专利信息资源检索的初级用户，只需要输入简单的专利主题词或关键词即可，为了缩小专利的查询范围，可以在发明专利、实用新型专利、外观设计专利和发明授权专利之间选择。同时可以只针对中国专利或世界专利进行检索。普通检索界面实例如图 12-38 所示。



图 12-38　专利搜索引擎的普通检索界面

6. 专利搜索引擎的高级检索

专利搜索引擎的高级检索一般包含熟悉应用以下一些专利检索项及其多种逻辑组合。

第一，进行专利国别筛选，以便于在 9600 万项专利文献中大大缩小其需求的地域范围，也可以组合在几个国家之间查询需要的特定领域专利。

第二，确定或设置需求结果的排序，包括主题相关度排序、申请日期的升序或降序排列，也可以设置依据专利公开日的升序或降序排列检索结果的输出。

第三，专利号码检索，包括应用专利的专利文献号、申请号和优先权进行查询。

第四，专利检索词位置限定，包括专利检索词出现在专利标题、专利摘要或全部正文内容中。

第五，依据专利日期查询，包括专利公开日期的具体时间（年、月、日）及其时间段或者专利申请日期的具体时间及其时间段。

第六，分类号检索，应用国际专利号（IPC）或欧洲专利分类号（ECLA）查询。国际专利分类号检索网址：http://www.wipo.int/ipcpub/#refresh=page；http://epub.sipo.gov.cn/ipc.jsp。

第七，专利权人/发明人检索，包括用专利权人的名称及其国别代码查询，也可以用发明人的名称及其国别代码查询。检索实例如图 12-39 所示。

12.4.5　国外大型专利数据库系统

1. Derwent Innovations Index（DII）

将德温特世界专利索引（WPI）和德温特专利引文索引（PCI）的内容整合在一起，采用

SOOPAT 新世界！

搜索 高级搜索 IPC搜索 引文搜索 专利族搜索

国家/地区: ☑所有专利（包含102个国家和地区，超过9600万专利文献）
☐ US美国 ☐ CN中国 ☐ EP欧专局 ☐ WO世知组织 ☐ JP日本 ☐ KR韩国 ☐ FR法国 ☐ GB英国 ☐ CH瑞士 ☐ SE瑞典
☐ NO挪威 ☐ FI芬兰 ☐ IT意大利 ☐ ES西班牙 ☐ NL荷兰 ☐ PL波兰 ☐ DK丹麦 ☐ BE比利时 ☐ AT奥地利 ☐ IN印度
☐ DE德国 ☐ DD东德 ☐ CA加拿大 ☐ AU澳大利亚 ☐ RU俄罗斯 ☐ SU苏联 ☐ ZA南非 ☐ BR巴西 ☐ AR阿根廷 ☐ MX墨西哥
☐ EG埃及 ☐ EG土耳其 ☐ IL以色列 ☐ MY马来西亚 ☐ ID印尼 ☐ SG新加坡 ☐ TW中国台湾 ☐ HK中国香港

排序: ◉相关度 ○申请日期降序 ○申请日期升序 ○公开日期降序 ○公开日期升序

SooPAT 搜索 SooPAT 分析

号码

文献号: (例如：US7701068 或 EP2008543)
申请号: (例如：EP20080011367 或 PCT/AU2007/000295 或 WO2007AU00295)
优先权: (例如：US20000603065 或 WO2001US40084)

常用

所有: (例如：sea)
标题: (例如：car)
摘要: (例如：car AND sea)

日期

文献公开日期:◉ (例如：20080501或 2010)
○ 到 (例如：20080501 到 2010)
专利申请日期:◉ (例如：20080501或 2010)
○ 到 (例如：20080501 到 2010)

分类

国际专利分类号（IPC）: (例如：G06F19/00)
欧洲专利分类号（ECLA）: (例如：F03G7/10)

专利权人/发明人

专利权人（申请人/受让人）: (例如：SUN MICROSYSTEMS INC)
专利权人国别代码: (例如：US)
发明人: (例如：WILSON WILLIAM)
发明人国别代码: (例如：EP)

SooPAT 搜索 SooPAT 分析

图 12-39 专利搜索引擎的高级检索界面

ISI Web of Knowledge 平台，通过学术论文和技术专利之间的相互引证的关系，建立了专利与文献之间的链接，可以检索到全球 40 多个专利机构授权的发明及其引用信息。DII 收录全球 40 多个专利机构的 1300 万条基本专利，3000 万项专利。每周增加 25 000 多个专利，分为 Chemical、Electrical & Electronic、Engineering 三部分。在检索结果全记录中，单击“Original(原始)”按钮，可浏览、下载专利说明书全文。

1) Derwent Innovations Index(DII)概述

Derwent Innovations Index 提供 Derwent 专业的专利信息加工技术，协助研究人员简捷有效地检索和利用专利信息，鸟瞰全球市场，全面掌握工程技术领域创新科技的动向与发展。Derwent Innovations Index 还同时提供了直接到专利全文电子版的连接，用户只需单击记录中“Original Document”就可以立刻链接到 Thomson Patent Store，获取专利申请书的全文电子版。Derwent Innovations Index 所链接的专利全文电子版，包括以下专利机构所公布的专利全文：USPTO(美国专利局，1963 年以来)；German Patent and Trademark Office(德国专利和商标局，1968 年以来)；ESP(欧洲专利局，EP-A 1978 年以来，EP-B 1980 以来)；WIPO(世界知识产权组织，1978 以来)；日本专利申请书第一页的英文翻译(2000 年以来)；其他许多国家，比如，奥地利、比利时、前东德、丹麦、法国、爱尔兰、意大利、卢森堡、荷兰、西班牙、瑞士、摩纳哥等。DII 数据库具有以下特点：增强的专利信息数据库；一条记录记载一项发明；人工标引以确保检索更一致性和精确性；用有限文字完整覆盖专利重要信息点；收录全球 48 个主要专利授权机构的专利文档，包括 2510 万件发明(同族)；用一个简单且结构化的记录来表述专利说明书中所有重要信息(To present all the significant information from a patent specification in a single highly structured record)。

图 12-40 简洁地表明了 DII 专利索引的特点：地域涵盖面广、内容描述简洁、人工信息标引准确、分类清晰、结构简单。

2) 同族专利

同族专利是基于同一优先权文件，在不同国家或地区以及地区间专利组织多次申请、多次公布或批准的内容相同或基本相同的“一组专利文献”，也就是同样内容的专利，在不同国家申请(同一专利内容与名称需要在不同国家申请并获得各自国家的专利保护需要)所构成的同一专利族类。

例如，检索美国的无效专利的同族专利(指已被专利局授权或公布的专利，经过一定的法律进程，失去专利权保护或自始至终未获得专利权的保护)US7097696B2(一种带有油分离和易更换阀的双筒空气干燥器)。见图 12-41～图 12-43。

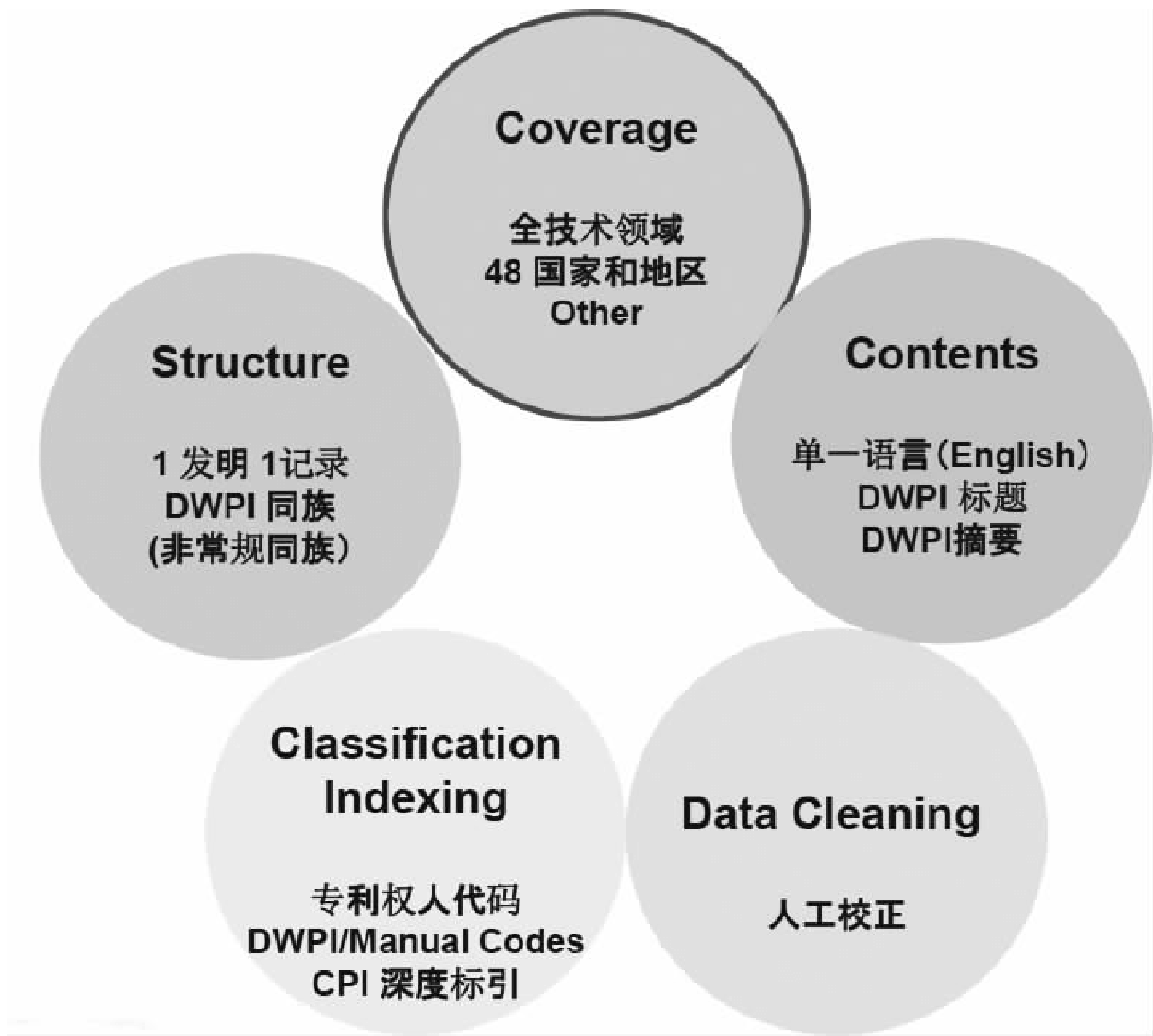


图 12-40　DII 专利索引的特点

记录视图: US7097696B2

添加至工作文件 | 标记记录 | 监控记录 | 下载 ▼ | 翻译 ▼ | 引证关系图 | 高亮显示 | 打印

完整浏览 跳转至: 著录项目 摘要 分类/索引 法律状态 同族专利 权利要求 说明书 引用 其他 自定义字段

+ 展开 施引专利 (10)

− 折叠 引用的专利 (14)　　作为检索结果查看

	公开号	公开日期	申请日期	发明人 DWPI	DWPI 专利权人/申请人	DWPI 标题	相关性	来源出版物
	US4108617A	1978-08-22	1977-02-07	FRANTZ V L	WHITE SALES CORP GRAHAM	Dual compressed air filter assembly for e.g. diesel locomotives which is valued etc. to avoid loss of input compressed air during switching from filter to purge	-	1 (Applicant)
	US5378266A	1995-01-03	1993-08-02	ELAMIN N A	ALLIED-SIGNAL INC	air system for industrial and automotive applications includes air drying subsystem which can be configured either for intermittent mode or continuous flow mode.	-	1 (Applicant)
	US4692175A	1987-09-08	1986-03-17	FRANTZ V L	ROANOKE COLLEGE	Two=stage pre-coalescer unit for contaminated compressed gas by coalescing loose oil and water in first coalescer and then aerosol in second combined coalescer	-	1 (Applicant)

图 12-41　DII 专利检索实例图(US7097696B2)

记录视图: US7097696B2

添加至工作文件 | 标记记录 | 监控记录 | 下载 ▼ | 翻译 ▼ | 引证关系图 | 高亮显示 | 打印

完整浏览 跳转至: 著录项目 摘要 分类/索引 法律状态 同族专利 权利要求 说明书 引用 其他 自定义字段

DPCI 引用 ? 同族专利级别

+ 展开 DPCI 施引专利 (15)

− 折叠 DPCI 引用的专利 (23)　　作为检索结果查看　　查找相关 DWPI 同族专利

公开号	入藏号	公开日期	申请日期	DWPI 同族专利成员	相关性	来源
US6280492B1	1996-269774	2001-08-28	1995-12-06	CN1921924B	-	0 (Examiner)
DWPI 标题: Flange for device for removing oil aerosols from air has rubber-coated metal plate which forms real between housing and fixing plate				KR1128881B1	-	0 (Examiner)
DWPI 专利权人/申请人: FILTERWERK MANN & HUMMEL GMBH (FILW-C)				US7097696B2	-	0 (Examiner)
DWPI 发明人: BINDER W; KELLER L; WOLF M				WO2005091783A2	A	0 (Examiner)
US4692175A	1987-270864	1987-09-08	1986-03-17	US7097696B2	-	0 (Examiner)
DWPI 标题: Two=stage pre-coalescer unit for contaminated compressed gas by coalescing loose oil and water in first coalescer and then aerosol in second combined coalescer						
DWPI 专利权人/申请人: ROANOKE COLLEGE (ROAN-N)						
DWPI 发明人: FRANTZ V L						

图 12-42　DII US7097696B2 同族专利检索实例图

DWPI 发明人: FRANTZ V L					
US5961698A	1999-264222	1999-10-05	1998-02-02	EP1718392A2	Y
DWPI 标题: Twin tower air dryer for cleaning and drying unpurified pressurised gas has manifold block to which separator and sump are mounted on one side and pair of desiccant containing canisters are mounted on other				US7097696B2	-
DWPI 专利权人/申请人: UOP INC (UNVO-C); WESTINGHOUSE AIR BRAKE CO (WESA-C)					
DWPI 发明人: DOSSAJI M R; FOSTER L L; GLENN T A; GURVIETCH S V; JONES C E; KAZAKIS M V; MCGEE C L; RYRIE B D; SHARMA S B; THOMAS G A					

图 12-43　DII US7097696B2 专利引用检索实例图

3）DII 专利检索视图

德温特专利检索系统遵循专业检索的普适性、用户检索的简捷性与友好性，分为表单检索、公开号检索和专家检索几种方式，实例如图 12-44 所示。

2．Espacenet

Espacenet 是欧洲专利局（EPO）的专利文献系统，可以免费检索 80 多个国家和地区的专利，其中大部分专利有全文。检索语言可以设置为英语、法语或德语。检索框每次可以接受最多 20 个检索词，多个检索词之间用空格隔开，用户可以检索 9000 万项专利文



图 12-44 DII 用户检索通用界面实例

献。普通检索常常也称为模糊检索、智能检索,它对于普通检索用户而言是简单高效的,不需要清楚查询词的位置是否在标题、摘要或正文内容中,也不需要清楚查询词是否为关键词、主题词等属性,但是检索结果的准确度不高,用户对检索结果的筛选与判别的难度和工作量都比较高。普通检索界面如图 12-45 所示,每次查询时,最多可以使用的检索词数量为 20(分类目录库的最大检索词为 10 个),但是各个检索词之间需要一个空格符或者逻辑与、或、非运算符(检索算子)进行分隔。



图 12-45 Espacenet 专利检索系统——普通检索界面

除了智能检索(或普通检索)之外,Espacenet 专利系统也提供高级检索功能,这对于熟练用户或专业人员而言是非常重要的,能够大大提高专利检索的查准率。总体上,Espacenet 高级检索包括了范围选择、专利题目、摘要、专利公开号、专利申请号、专利日期、专利权人等十多项高级组配检索功能,实例图如图 12-46 所示。

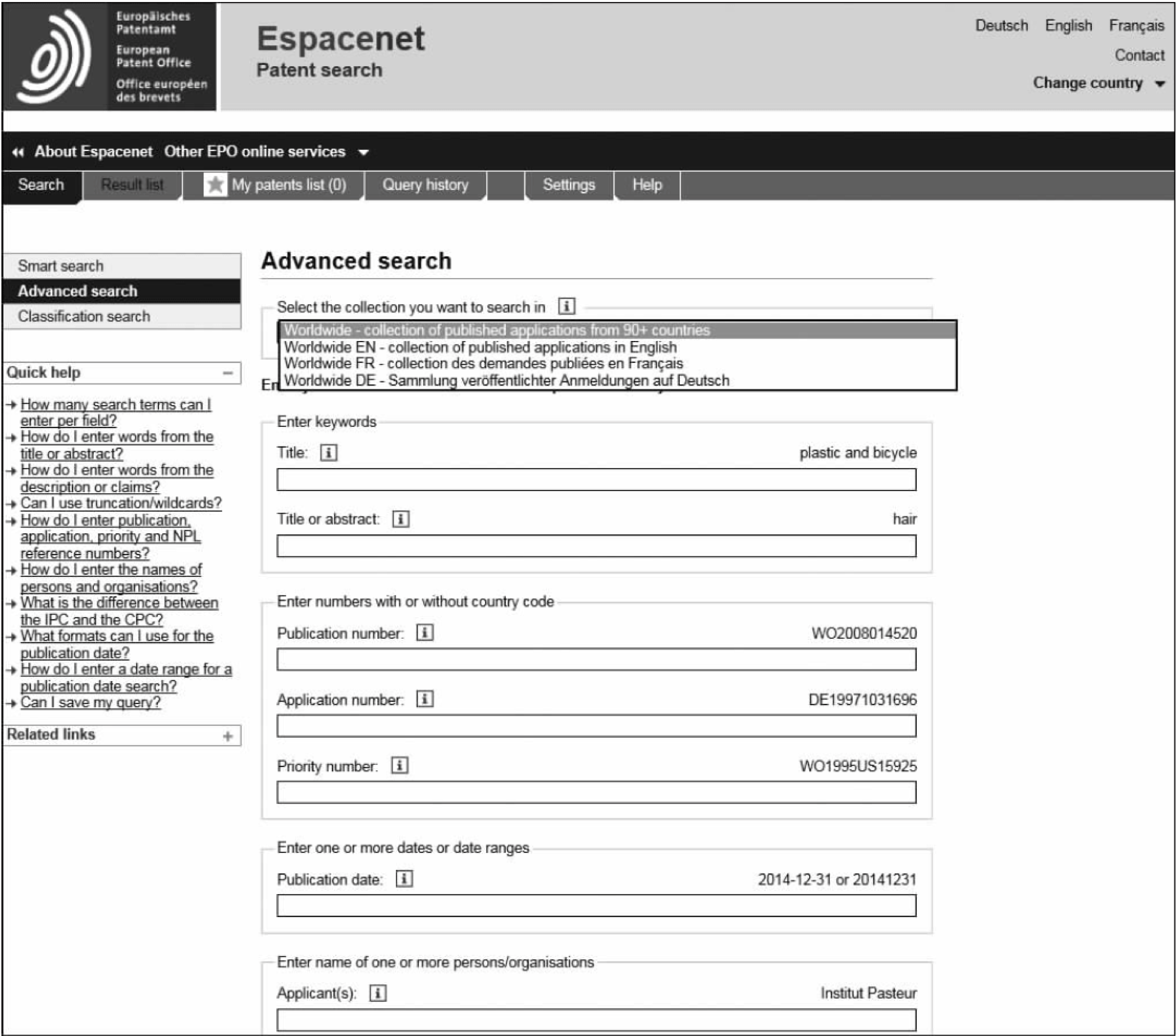


图 12-46　Espacenet 专利检索系统——高级检索界面

3. USPTO Patent Full-Text and Image Database

USPTO 即美国专利与商标局(United States Patent and Trademark Office)的简称,美国专商局在促进有效与均衡的全球知识产权保护方面一直处于领先地位。美国专商局的任务:利用其能力强、多元化的人才队伍,通过提供及时、高质量的专利与商标审查、指

导国内与国际知识产权政策、向全球提供知识产权信息和教育等工作，在本国和世界范围内促进创新、竞争力和经济增长。在系统中可以检索1790年以来的所有美国专利，可以在线浏览全文(tif 文件)，但需要下载浏览器 alternatiff。

USPTO 专利快速检索主要容纳两个检索词，以及确定检索词的范围关系(包括专利名称、摘要、申请系列号、分类号、专利权人、发明人、申请国别等近30项范围限定)。其检索界面见图12-47。

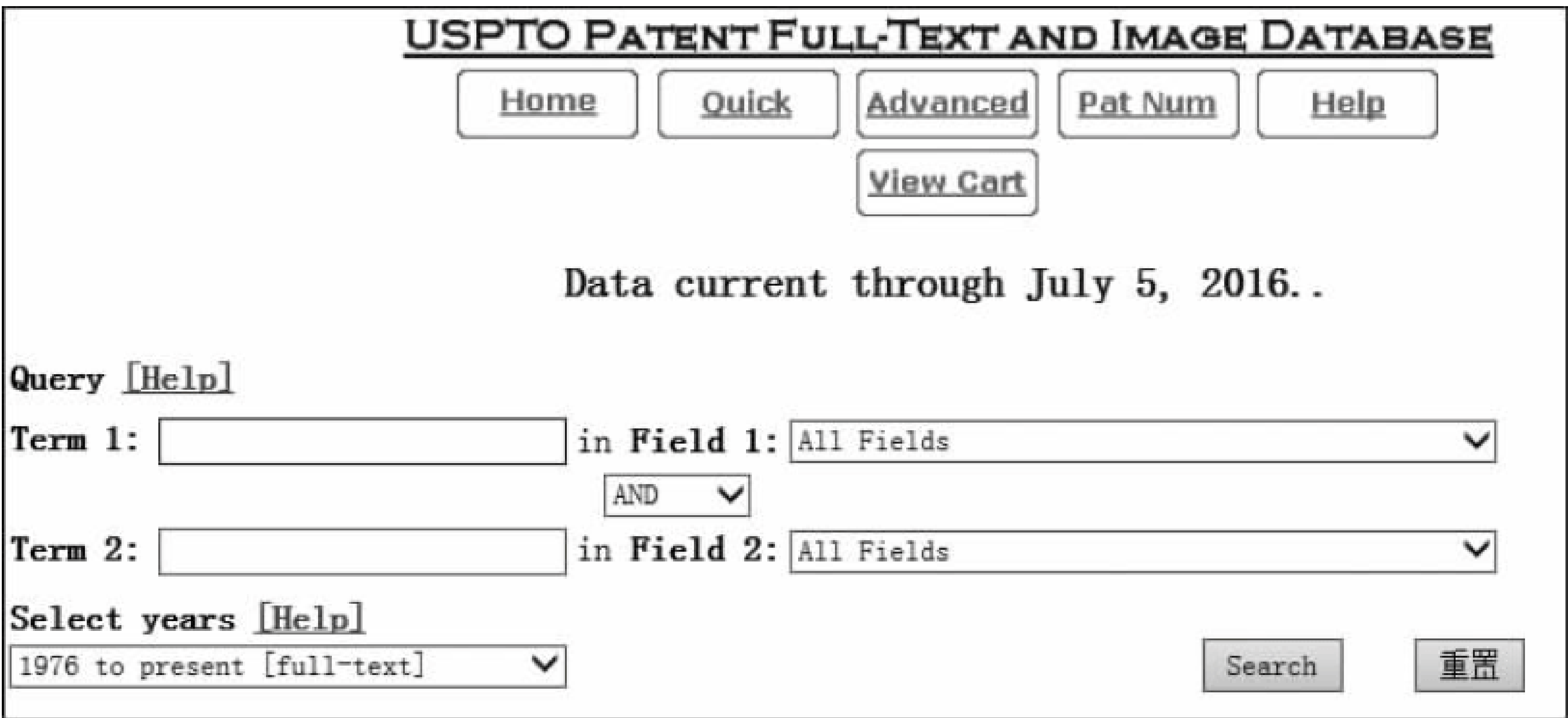


图12-47　USPTO 专利快速检索界面

USPTO 专利高级检索主要是专利的限定与逻辑组合检索，例如，ttl/(tennis and (racquet or racket))，isd/1/8/2002 and motorcycle，in/newmar-julie。其检索界面见图12-48。

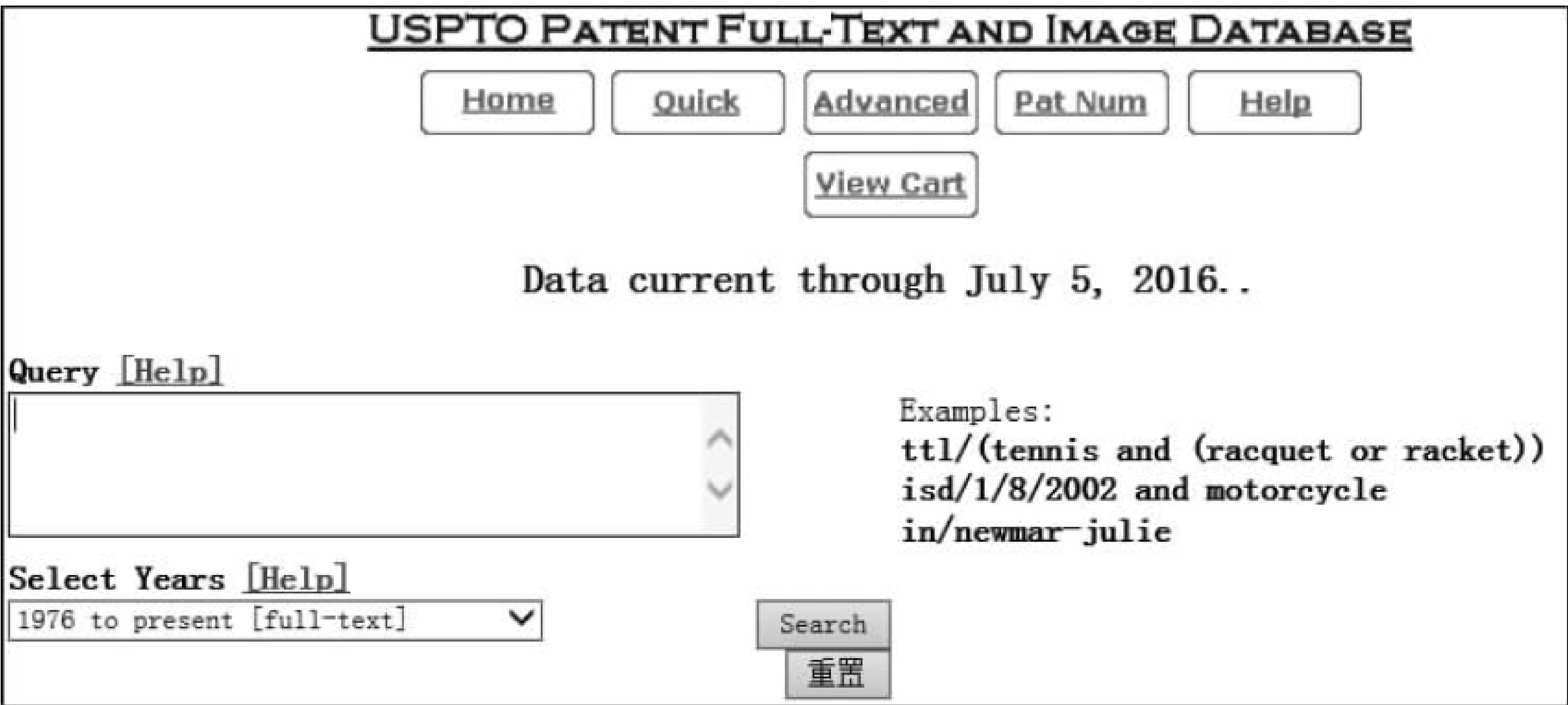


图12-48　USPTO 专利高级检索界面

4. PATENTSCOPE

世界知识产权组织(WIPO)是关于知识产权服务、政策、合作与信息的全球组织，是

一个自筹资金的联合国机构，有 188 个成员国，其使命是领导发展兼顾各方利益的有效国际知识产权制度，让创新和创造惠及每个人。该组织的任务、领导机构和工作程序载于《WIPO 公约》。

PATENTSCOPE 世界知识产权组织的免费专利数据库(https://patentscope. wipo. int)，包括多语言检索界面，系统包括 290 万国际专利申请(PCT)和 5700 万地区及国家汇编专利文献。该数据库提供四种专利检索方式：简单检索(simple search)、高级检索(advanced search)、字段组合检索(structured search)、多语种交叉扩展检索(cross lingual expansion search)，还可以浏览每周公布的专利文献等检索，大部分专利有全文内容。PATENTSCOPE 的中文检索界面见图 12-49。



图 12-49　PATENTSCOPE 的中文检索界面

5. 日本专利查询——特许、实用新案公报专利数据库

该数据库可以检索日本专利，并可看到部分日本专利说明书全文。

6. 加拿大知识产权局专利数据库

加拿大知识产权局专利数据库(Canadian Intellectual Property Office Canadian Patents Database)可以检索加拿大专利(http://www. ic. gc. ca/opic-cipo/cpd/eng)，检索方式有基本检索(basic search)、代码检索(专利号检索，number search)、布尔检索(boolean search)和高级检索(advanced search)。加拿大专利数据库布尔检索界面实例见图 12-50。

图 12-50 中表明对专利检索词或检索短语进行布尔组合检索时，可以对检索词、布尔算子、文本域进行交互控制，同时也可以对专利文献的状态、文献类别以及专利的各种日期数据(申请日、公开日等)进行交互操作。

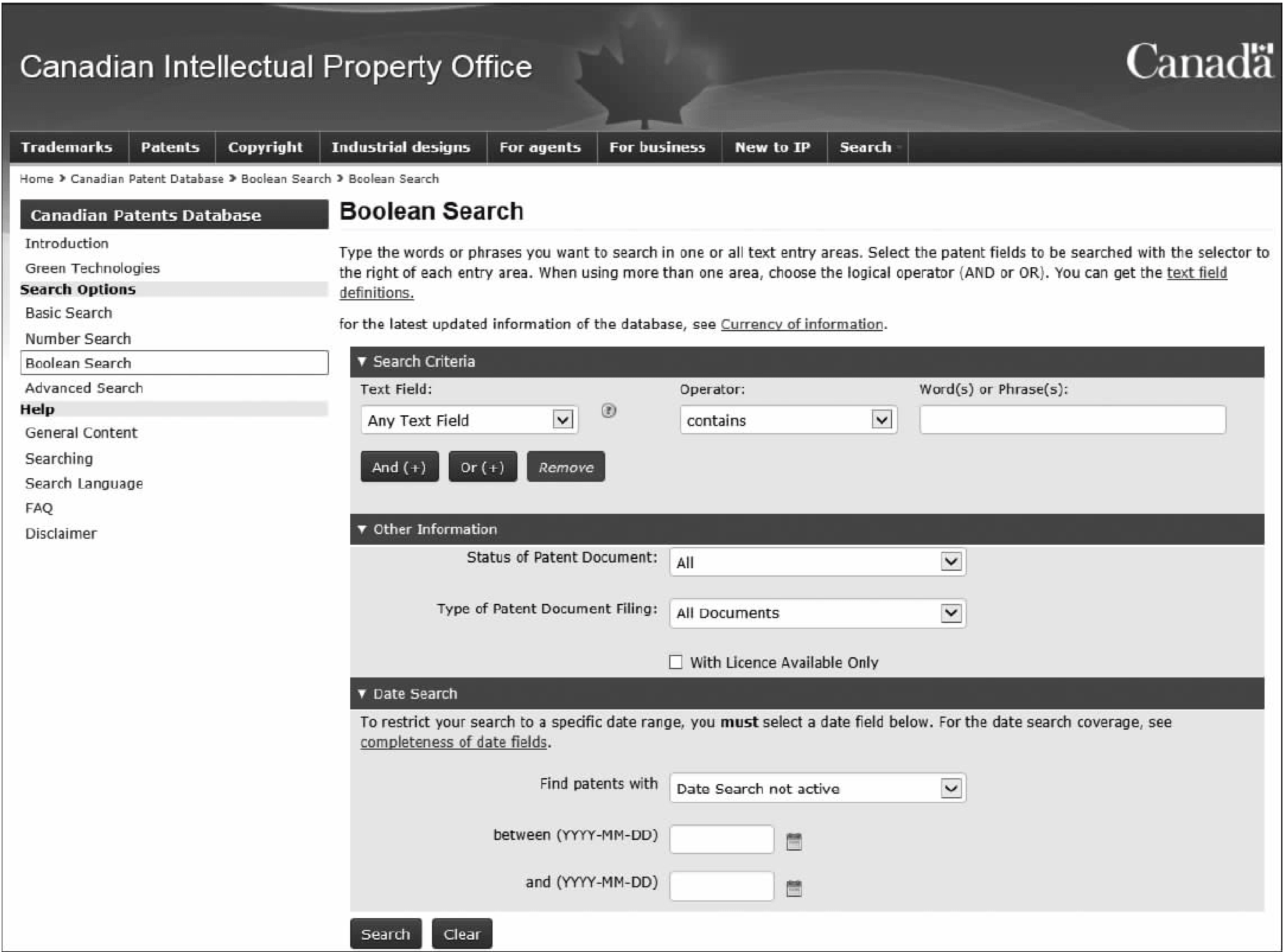


图 12-50　加拿大专利数据库布尔检索界面实例

7. LexisNexis

LexisNexis 是世界著名的数据库，全球许多著名法学院、法律事务所、高科技公司的法务部门都在使用该数据库。该数据库连接至 40 亿个文件、11 439 个数据库以及 36 000 个来源，资料每日更新。其中的专利数据库收录 1980 年以来的欧、美、日的专利全文，也包括关于专利法律研究的信息内容，通过其中 Patent Law 专栏中的 Patent 数据库，可以检索并在线浏览专利全文，包括美国专利、欧洲专利、英国专利、世界专利、日本专利和通过 PCT 申请的专利。数据库以分类浏览的方式，单击具体的类目进行查看，里面包含一些图标代表着不同的含义。

美国 LEXIS-NEXIS 公司创始于 1973 年，其数据库内容涉及新闻、法律、政府出版物、商业信息及社会信息等，其中法规法律方面的数据库是 LEXIS-NEXIS 的特色信息源，具有非常大的影响力，尤其在法律业界具有很高知名度。LexisNexis Academic 是

Lexis-Nexis 数据库产品中,面向大学和学术研究设计的数据库。共选自 5300 种出版物的内容,主要包括以下几个方面的主题:综合性新闻;公司商业信息;政府规章、政治新闻、法律研究 ;医学、保健信息;参考性资料数据库。

LexisNexis 系统的一般检索可以输入关键词、标题、作者或 ISBN 号进行模糊查询。其检索界面见图 12-51。

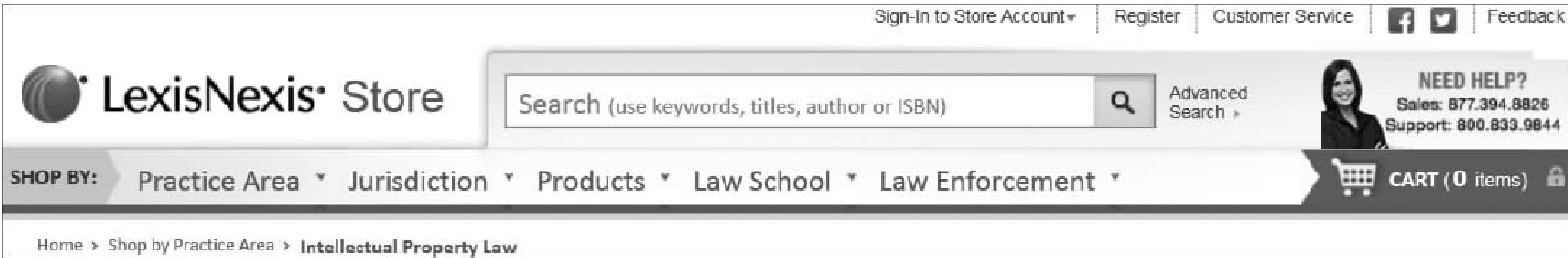


图 12-51　LexisNexis 系统的一般检索界面

12.4.6　国内专利资源数据库系统检索

1. 国家知识产权局专利检索

(1) 概述。国家知识产权局(State Intellectual Property Office)1980 年经国务院批准成立中国专利局,1998 年更名为国家知识产权局,国家知识产权局对专利申请的受理、审查、复审、授权以及对无效宣告请求的审查业务委托国家知识产权局专利局承担。

(2) 检索功能与专利信息范围。检索功能包括常规检索、表格检索、药物专题检索、检索历史、检索结果浏览、文献浏览、批量下载等。分析功能:快速分析、定制分析、高级分析、生成分析报告等。专利数据范围:收录了 103 个国家、地区和组织的专利数据以及引文、同族、法律状态等数据信息,其中涵盖了中国、美国、英国、法国、德国、俄罗斯、欧洲专利局和世界知识产权组织等专利数据资源。

(3) 一般检索。一般检索通过自动识别、检索要素、申请号、申请人和发明人等开展快速检索。其检索界面见图 12-52。



图 12-52　国家知识产权局专利检索系统的一般检索界面

自动识别：即自动识别逻辑检索，检索词之间分隔符、时间格式、小括号与双引号等自动识别。见图 12-53。



图 12-53　国家专利局专利系统“自动识别的检索含义”

检索要素：在专利标题、专利摘要、权利要求和分类号中同时检索，也可以加双引号查询。见图 12-54。



图 12-54　国家专利局专利系统“检索要素的检索含义”

申请号：申请号的检索含义包括申请号格式应用、自动去掉校验位、支持模糊匹配和截词符等。见图 12-55。



图 12-55　国家专利局专利系统“申请号的检索含义”

(4)表格检索。以表格的形式将专利检索项进行排列，便于用户进行精确匹配检索，同时支持命令编辑和复杂检索式的逻辑构造。例如“摘要＝(computer)or 申请日＝20151013:20160723”，在本系统中，表格式检索的含义与作用等同于高级检索或专业检索。表格式检索界面见图 12-56。

2. 中国专利信息中心专利之星检索系统

中国专利信息中心成立于 1988 年，是国家知识产权局直属事业单位、国家级大型专利信息服务机构，拥有国家知识产权局赋予的专利数据库管理权、使用权和综合服务经营权。拥有完整稳定的专利数据资源、多功能综合性专利检索系统、企业创新专家支持平台，承接政府、机构、企业、公众等的专利数据处理、数据提供、检索咨询、定制化开发等业务。

(1) 表格检索，提供各个专利项精确匹配的检索交互界面，包括命令行检索(以命令行构造用户需要的复杂逻辑检索表达式)。见图 12-57。

图 12-56　国家专利局专利系统表格式检索界面

图 12-57　中国专利之星检索系统的表格式实例

(2) 专家检索，提供专业性强、检索精度高的专利检索服务。见图 12-58。

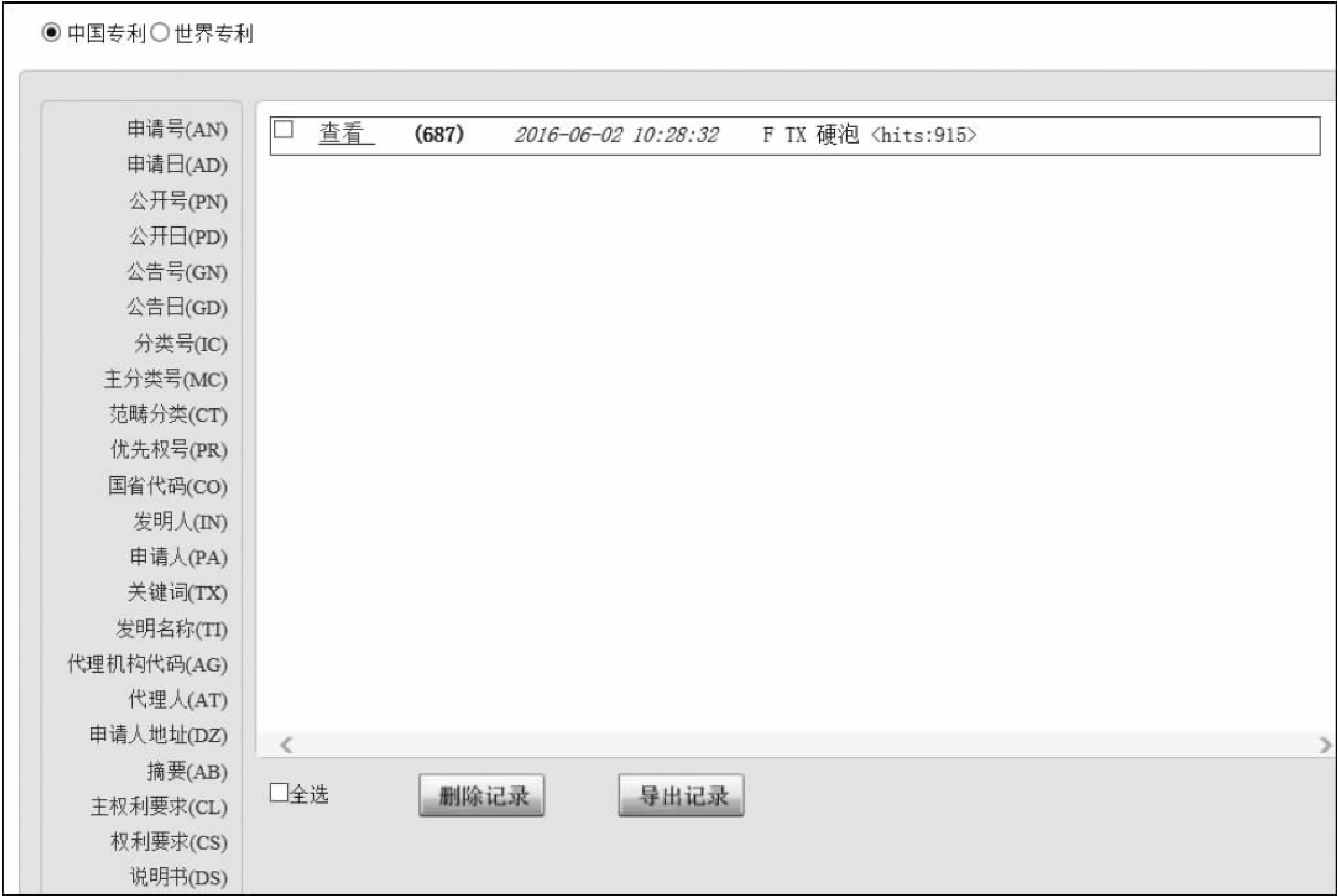


图 12-58　中国专利之星检索系统的专家检索模块界面

(3) 专利之星主要检索模块，包括智能检索、表格检索、专家检索和法律状态检索四大类检索，以及分类导航、专利预警、专题数据库和中外专利文献的机器翻译等辅助检索功能，可检索中国专利和世界专利，并可下载 PDF 全文，部分功能注册后才可使用。见图 12-59。



图 12-59　中国专利之星检索系统的主要检索功能模块

3. 万方中外专利数据库

万方中外专利数据库(wanfang patent database, WFPD)，收录始于 1985 年，4500 余

万项中外专利，年增 25 万条。

万方中外专利数据库的专门检索工具为 PATENTOOL，专利文献来源于 11 国（中国、美国、澳大利亚、加拿大、瑞士、德国、法国、英国、日本、韩国、俄罗斯）和两组织（世界专利组织、欧洲专利局）。一般检索首先在选择专利类型（全部类型、发明专利、实用新型和外观设计）的基础上，输入检索的关键词即可。其一般检索界面见图 12-60。为了提高检索结果的准确率，使用其高级检索是明智的。



图 12-60　万方中外专利数据库“一般检索界面”

高级检索可以细化专利检索内容项，包括申请号、公开号、名称、摘要、申请日、公开日等十多项的逻辑组合，以提高查准率。高级检索界面见图 12-61。

4. Patent Cloud（专利云检索网）

Patent Cloud 是由富士康公司开发的专利文献资源检索系统，Patent Cloud 包含中国大陆、中国台湾、美国、韩国、日本、WO、欧洲专利信息，累计 4700 万篇专利。Patent Cloud 支持简体中文、繁体中文、英文等多种语言版本，可以提供专利家族信息、引证信息等，保存 PDF 格式全文。

（1）检索语法详见相关参照，其中词组（或两词）如未加双引号，系统默认两词之间逻辑关系为“or”。

（2）美国、中国台湾专利可以进行全文检索。

（3）支持 IE9.0 以上版本，支持 Google 浏览器及其他浏览器。

（4）多元检索方式快速查询专利资料，并能实时产生分析报表，让用户实时精准探索专利大数据，同时辅以多种实用工具，提供用户专属的线上专利工作平台。

图 12-61　万方中外专利数据库“高级检索界面”

（5）同步显示两篇专利全文，帮助用户比较内容异同。

5. 中国专利/海外专利全文数据库（知网版）

《中国专利全文数据库（知网版）》包含发明专利、实用新型专利、外观设计专利三个子库，准确地反映了中国最新的专利发明。专利相关的文献、成果等信息来源于 CNKI 各大数据库，可以通过申请号、申请日、公开号、公开日、专利名称、摘要、分类号、申请人、发明人、优先权等检索项进行检索，并一次性下载专利说明书全文。按照专利种类分为发明专利、外观设计和实用新型三个类型，其中发明专利和实用新型采用国际专利分类法（IPC 分类）和 CNKI 168 学科分类，外观设计采用国际外观设计分类和 CNKI 168 学科分类。收录从 1985 年至今的中国专利。截止到 2016 年 5 月，《中国专利全文数据库》共计收录专利 1000 多万条。

与通常的专利数据库相比，《中国专利全文数据库》（知网版）每条专利的知网节集成了与该专利相关的最新文献、科技成果、标准等信息，可以完整地展现该专利产生的背景、最新发展动态、相关领域的发展趋势，可以浏览发明人与发明机构更多的论述以及在各种

出版物上发表的文献。

在用户检索方面，有初级检索、高级检索和专业检索三种，图 12-62 是高级检索界面实例。



图 12-62　中国知网中国专利数据库“高级检索界面”

12.5　标准信息资源检索

12.5.1　标准信息资源的概念与特点

狭义的标准信息资源是指按规定程序制定，经公认权威机构或主管机关批准的一整套在特定领域内必须执行的规格、规则、技术要求等规范性文献资料，简称标准。标准是大学生获取的一种重要学习与参考资源类型。广义的标准指与标准化工作有关的一切信息资源，包括标准形成过程中的各种档案、宣传推广标准的手册及其出版物、揭示报道标准文献信息的目录、索引等。国外标准信息资源经常使用的名称有标准(standard)、规格(specification)、公报(bulletin)、建议(recommendation)、法规(code)、手册(handbook)、规则(rules instruction)和工艺(practice)等。在公元前 1500 年的古埃及纸草文献中即有关于医药处方计量方法的标准，是现在最早的标准。通常认为，现代标准文献资源产生于 20 世纪初。1901 年英国成立了第 1 个全国性标准机构，同年世界上第 1 批国家标准问世。此后许多发达国家相继建立全国性标准化机构出版各自的标准，其中影响较大的有

美、英、德、法、日、俄等国家。1906 年成立的国际电工委员会(International Electrotechnical Commission,IEC)和 1947 年 2 月成立的国际标准化组织(International Organization for Standardization,ISO)是两个最重要的国际标准机构(1947 年将 IEO 并入 ISC,但在技术、财政、名称及工作程序上仍保持独立性)。随着标准化事业的发展,标准文献资源也急骤增长。世界各国的各类标准文献连同相关的会议文件、技术报告等,数量更是高达数千种。1956 年,我国设立国家标准局,1957 年 8 月加入 IEC,并颁布了第 1 批国家标准。1978 年 5 月成立国家标准总局,1988 年我国组建国家技术监督局。1989 年 4 月 1 日《中华人民共和国标准化法》实施。

标准按使用范围划分有国际标准、区域标准、国家标准、专业标准、地方标准、企业标准;按内容划分有基础标准(一般包括名词术语、符号、代号、机械制图、公差与配合等)、产品标准、辅助产品标准(工具、模具、量具、夹具等)、原材料标准、方法标准(包括工艺要求、过程、要素、工艺说明等);按成熟程度划分有法定标准、推荐标准、试行标准、标准草案。国际标准由国际标准化组织(ISO)理事会审查,ISO 理事会接纳国际标准并由中央秘书处颁布;国家标准在中国由国务院标准化行政主管部门制定,行业标准由国务院有关行政主管部门制定,企业生产的产品没有国家标准和行业标准的,应当制定企业标准,作为组织生产的依据,并报有关部门备案。

12.5.2　标准信息资源的分类

1.《中国标准文献分类法》

《中国标准文献分类法》(Chinese Classification for Standards,CCS)由国家技术监督局编辑,中国标准文献出版社 1989 年出版。《中国标准文献分类法》的类目设置以专业划分为主,适当结合科学分类。序列采取从总到分,从一般到具体的逻辑系统。该分类法采用二级分类,一级类目设置主要以专业划分为主,二级类目设置采取非严格等级制的分类方法。一级分类由 24 个大类组成,每个大类有 100 个二级类目;一级分类由单个拉丁字母组成,二级分类由双数字组成。

2.《国际标准分类法》

《国际标准分类法》(International Classification Standards,ICS)是由国际标准化组织 1991 年组织编制的,主要用于国际标准、区域标准和国家标准以及其他标准文献的分类。国际标准分类法的推广应用,有利于标准信息资源分类的协调统一,促进国际间标准文献的交换与传播。ICS 采用三级数字编号,第 1 级由 41 个大类组成,第 2 级为 387 个二级类,第 3 级为 789 个小类。第 1 级和第 3 级用双位数表示,第 2 级用三位数表示,各级类

目之间以圆点相隔。如71.040.50代表物理化学分析方法。

12.5.3 美英等国标准信息资源检索

1. 美国标准及其检索

美国国家标准(American National Standards)创建于1918年,由美国国家标准学会(American National Standards Institute,ANSI)负责制定。ANSI标准采用字母与数字相结合的混合标记分类法。用1个字母标记1个大类,用数字表示大类目下的小类。1984年前,ANSI共分24个大类;1985年以后,按《美国国家标准协会目录》分为17大类。

(1)《美国国家标准学会目录》。该目录由美国国家标准学会编辑出版,每年出版一次,是美国标准的主要检索工具书。目录中列举了现行美国国家标准,内容包括两个主要部分,即"主题目录"(listing by subject)和"标准序号目录"(listing by designation)。在各条目下列出标准主要内容、标准制定机构名称代码和价格,可以从主题和序号途径查找美国国家标准。

(2)《美国试验与材料协会标准年鉴》(American Book of ASTM Standards)。该年鉴由美国试验与材料协会(American Society for Testing and Material,ASTM)编辑出版,是查找该协会制定的标准的主要检索工具,每年出版1次。该年鉴分16个部分,66卷,按专业分类。《ATSM标准年鉴》中可供检索用的主要有两个栏目:一个是主题索引(subject index),是年鉴中综合主题索引;另一个是字母序号表(a1phanumeric list),在此表中,按字母及序号的次序列出了全部ASTM现行标准和暂行标准。ASTM系统(http://compass.astm.org/)需要注册后使用。

(3)《联邦规格标准和商品说明书索引》(*Index Federal Specification, Standard Commercial Item Dessification*)。该索引由美国总务管理局(General Services Administration)编辑出版,每年出一版,是查找美国联邦规格和标准的主要检索工具。内容主要有三部分:"字顺一览表"、"序号一览表"、"联邦供应分类一览表"。可以按字顺、序号和分类三种途径查找到该组织制定的标准和标题标准号码、合格产品目录、联邦供应分类、主编单位、日期和价格。图12-63是"美国国家标准学会(American National Standards Institute,ANSI,http://webstore.ansi.org/sitelicense.aspx)"的一般检索界面,输入需要检索的标准的关键词即可。

美国ANSI检索系统可以选择需要检索的范围(例如ISO标准、IEC标准等),查询时可以输入标准号和关键词。见图12-64。



图 12-63 美国国家标准学会的一般检索界面

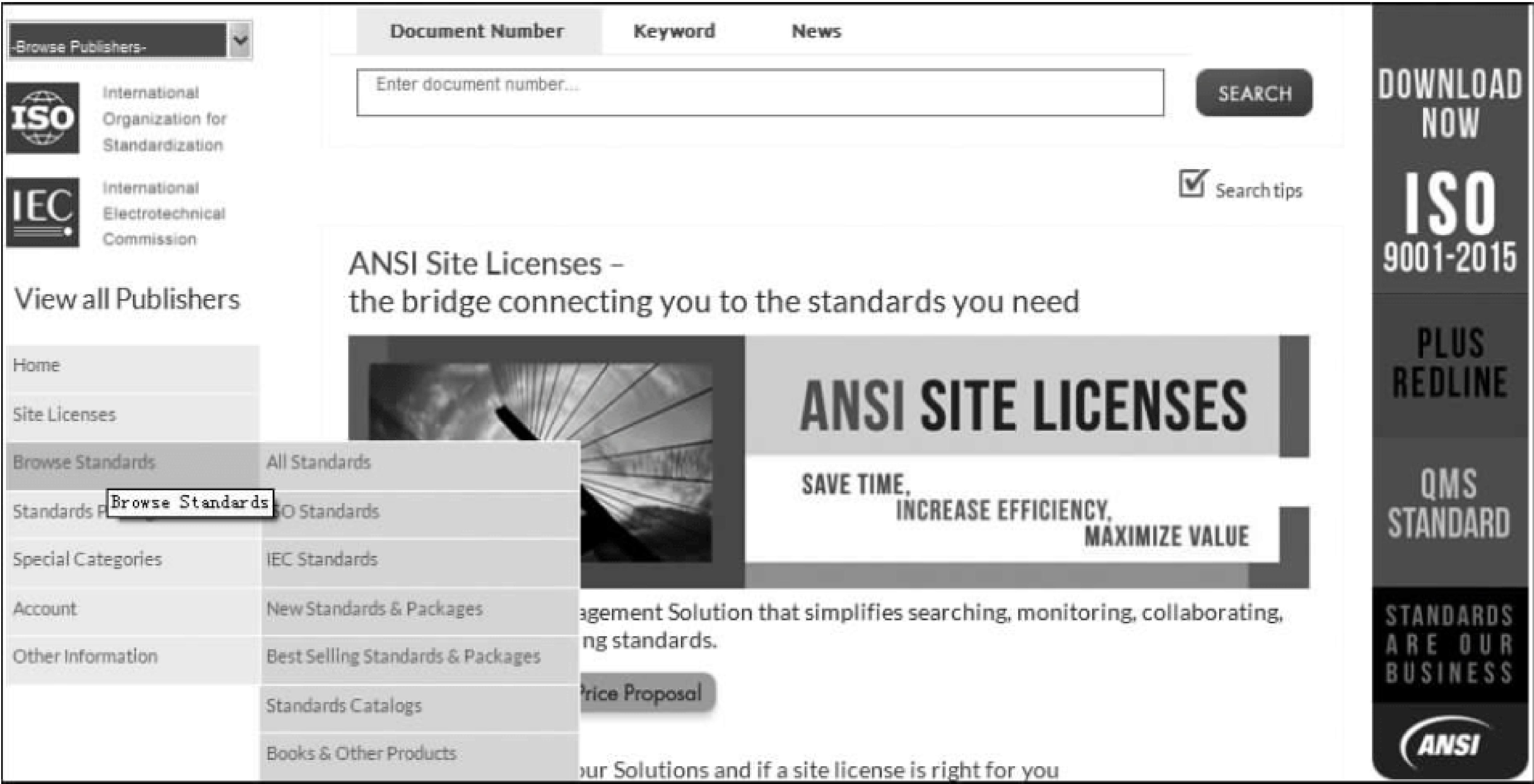


图 12-64 美国 ANSI 检索系统界面

2. 日本国家标准及其检索

日本工业标准(HS)由成立于1949年的日本工业标准调查会(Japanese Industrial Standards Committee,JISC)负责制定。该调查会下设29个部会,2000多个专门委员会。目前,现行一万多件标准,每隔5年审议1次。日本工业标准为国家级标准,除药品、食品及其他农林产品另行制定专门技术规范或标准外,涉及各个工业领域,内容包括技术发明及符号,工业产品的形状、质量指数及性能,试验、分析与测量,设计、生产、使用及包装运输等方法。检索日本国家标准的检索工具主要有以下两种。

(1)《日本工业标准总目录》(HS总目录)。该目录由日本标准协会编辑出版,每年出

一版，报道收集到当年3月份为止的全部日本工业标准。主要内容分为两部分：第1部分是“JIS总目录”，为专业分类下的标准序号索引；第2部分为主题索引。同时还附有ISO和IEC技术委员会的名称表、主要国外标准组织一览表及HS和日本事业标准制定单位一览表等。该目录提供有分类途径和主题途径。分类途径：使用分类目录查找，先确定课题所属的部类和小类，并按所指页次逐一查找，即能获得所需标准。主题途径：索引按日文字母顺序列出一级和二级主题词，并在其后著录相关标准的标准号。

(2)《日本工业标准年鉴》(*JIS Yearbook*)。此年鉴系英文版的日本工业标准目录。此外，还有《标准化杂志》、《日本工业标准手册》等多种检索工具及相关网站可以使用。

3. 德国国家标准及其检索

现行德国国家标准采用原联邦德国标准，由德国标准学会负责制定。该组织成立于1917年，原名为德意志工业标准委员会，1975年改为现名。联邦德国标准学会是一个注册的民间组织团体，1975年与联邦德国政府签署协议，政府承认该学会是德国标准化主管单位，具有法定资格，该学会制定的标准为联邦德国国家标准，目前该协会标准2万件。

检索联邦德国标准的工具主要有以下两种。

(1)《联邦德国标准学会技术标准目录》。该目录每年出版1次，报道到上一年年底为止的现行标准。内容除了联邦德国标准外，还列出联邦德国工程师协会、联邦德国航空标准组织、联邦德国国际防御装备标准组织的标准。目录内容分为两部分：一部分是“国际十进位分委法的主题集”和作为检索之用的主要部分的“主题集”，实质为国际十进位分类目录；另一部分是“数字索引”、“德文主题索引”和“英文主题索引”(English Index of Subject)。该目录提供有分类、序号和主题途径。

(2)《联邦德国标准化通报》由联邦德国标准学会编辑出版，月刊。报道标准化论文和有关国内外标准化新闻以及新颁布标准等。

4. 英国国家标准及其检索

英国国家标准的主体是英国标准(British Standard，BS)，由创建于1901年的英国标准学会(British Standards Institution，BSI)负责制定。BSI分标准、质量保证、信息服务与市场、公共事务、财务计算机管理、人事财产等多个部门，下设近千个技术委员会。英国标准在世界上有较大影响，因为英国是标准化先进国家之一，并为英联邦国家采用，所以英国标准受到国际上的重视。英国标准5年复审1次，现行标准1万多件。英国国家标准及有关出版物主要有下列几种类型：一般标准(BS)、实用规范(CP)、手册和专辑(Handbook，PD)等。检索英国标准的主要工具有以下三种：

(1)《英国标准学会目录》(*British Catalogue*)。该目录由英国标准学会按年度编辑

发行。

(2)《英国标准学会通报》(*BSI News*)。月刊,1946 年创刊,由英国标准学会编辑出版,报道标准化理论、国内标准以及 ISO、IEC 标准的动态。

(3)《英国标准学会年报》(*The BSI Annual Report*)。由英国标准学会编辑出版,报道英国标准学会、ISO 及 IEC 各委员会的工作成果。此外,还有《英国标准年鉴》(*British Standards Yearbook*)、中文版的《英国标准目录》等检索工具以及相关的网站检索系统可供使用,英国标准学会(BSI)检索系统如图 12-65 所示。

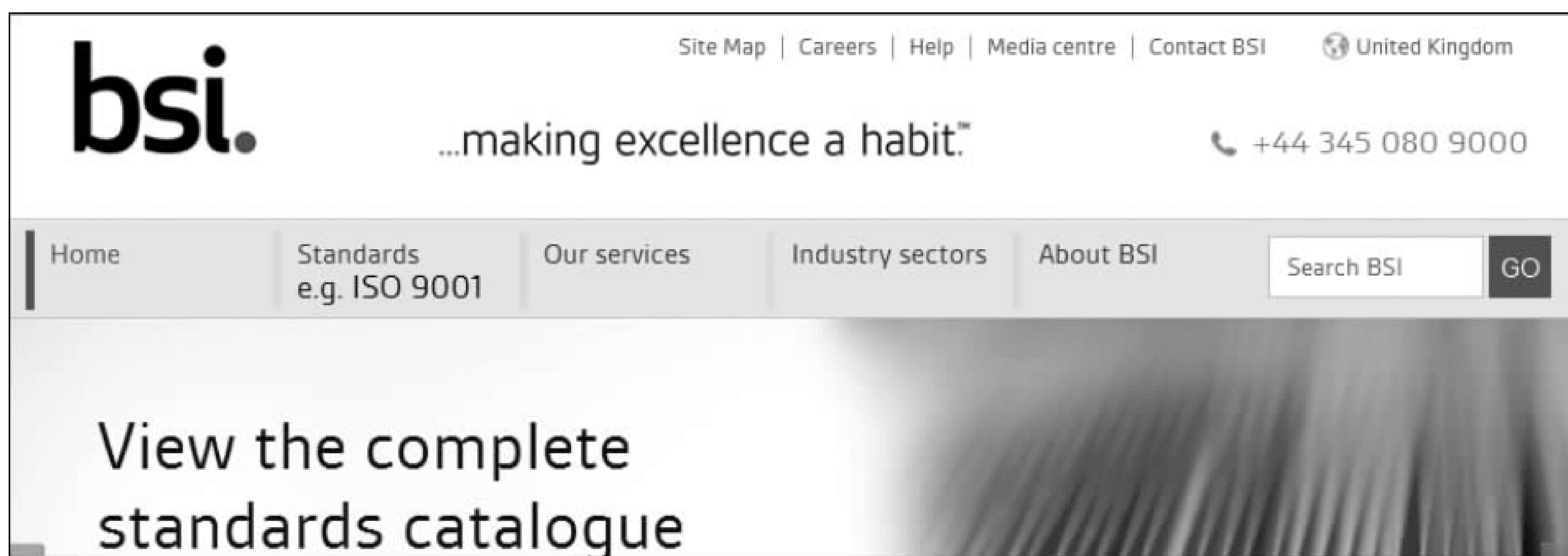


图 12-65　英国标准学会(BSI)检索系统

12.5.4　中文标准信息资源检索

1. 万方标准检索系统

综合了由国家技术监督局、建设部情报所、建材研究院等单位提供的相关行业的各类标准题录。截止到 2016 年该系统包括中国标准、国际标准以及各国标准等近 43 万项。

(1) 万方标准的分类检索。万方依据行业的不同把标准分为综合、农业、医药、矿业、航空等一级大类 24 个,二级分类有针织、棉纺织、铁路通信等 217 种,便于用户进行分类检索,实例如图 12-66 所示。

(2) 万方标准的高级检索。在标准类型方面可以选择中国国家标准、中国行业标准、国际标准化组织标准、欧洲标准、美国标准等,同时可以细化标准号、标准名称、关键词、国别、发布单位、起草单位等内容,以达到精确检索的目的。高级检索界面见图 12-67。

(3) 万方标准的专业检索。标准的专业检索就是要构造专业 CQL,即构造专业的逻辑检索表达式,实现高查准率的目的。专业检索界面见图 12-68。

机械

机械综合	通用零部件	加工工艺
工艺装备	金属切削机床	通用加工工艺
通用机械与设备	活塞式内燃机与其他动力设备	

电工

电工综合	电工材料和通用零件	旋转电机
低压电器	输变电设备	发电用动力设备
电气设备与器具	电气照明	电源
电工生产设备		

电子元器件与信息技术

电子元器件与信息技术综合	电子元件	电真空器件
半导体分立器件	光电子器件	微电路
计算机	信息处理技术	电子测量与仪器
电子设备专用材料、零件、结构件	电子工业生产设备	

通信、广播

通信、广播综合	通信网	通信设备
雷达、导航、遥控、遥测、天线	广播、电视网	广播、电视设备
邮政	通信、广播设备生产机械	

图 12-66　万方标准检索系统的标准分类检索部分目录实例

万方数据
WANFANG DATA
知识服务平台

» 检索首页 > 标准高级检索

高级检索　经典检索　专业检索

高级检索

标准类型：全部
- 全部
- 中国国家标准
- 中国行业标准
- 国际标准化组织标准
- 国际电工委员会标准
- 欧洲标准
- 英国标准化学会标准
- 法国标准化学会标准
- 德国标准化学会标准
- 日本工业标准调查会标准
- 美国国家标准学会标准
- 美国机械工程师协会标准
- 美国材料试验协会标准
- 美国电气及电子工程师学会标准
- 美国保险商实验室标准

标准编号：
任意字段：
标题：
关键词：
国别：
发布单位：
起草单位：
中国标准分类号：
国际标准分类号：
发布日期：
实施日期：　-　年
确认日期：　-　年
废止日期：　-　年
排序：◉相关度优先 ○发布日期优先
每页显示：10
检索

图 12-67　万方标准检索系统的高级检索界面



图 12-68　万方标准检索系统的专业检索界面

检索表达式使用[CQL 检索语言]，含有空格或其他特殊字符的单个检索词用引号("")括起来，多个检索词之间根据逻辑关系使用"and"或"or"连接。提供检索的字段：标准编号 StanCode、标准名称 Title、发布单位 IssueComp、发布日期 IssueDate、中国标准分类号 ChClass、关键词 Keywords、国别代码 StateCode。可排序字段：发布日期 IssueDate、相关度 Relevance。例如，加工 or IssueComp＝SBTS、Title All"电子政务"、中国标准 and Keywords＝食品等。

2. CNKI(中国知网)标准检索系统

《国家标准全文数据库》收录了由中国标准出版社出版的、国家标准化管理委员会发布的所有国家标准，占国家标准总量的 90％以上。标准的内容来源于中国标准出版社，相关的文献、专利、成果等信息来源于 CNKI 各大数据库。可以通过标准号、中文标准名称、起草单位、起草人、采用标准号、发布日期、中国标准分类号、国际标准分类号等检索项进行检索。标准的收录年限为 1950 年至今。

(1) CNKI 标准的初级检索。初级检索只需要确认检索(标准号、起草单位、实施日期等)后，输入检索关键词即可，用户也可再细化一些标准产生的时间段、匹配关系(模糊或精确)、结果相关度排序等内容。初级检索界面见图 12-69。

(2) CNKI 标准的高级检索。可以用逻辑与、逻辑或、逻辑非对标准号、起草单位、出

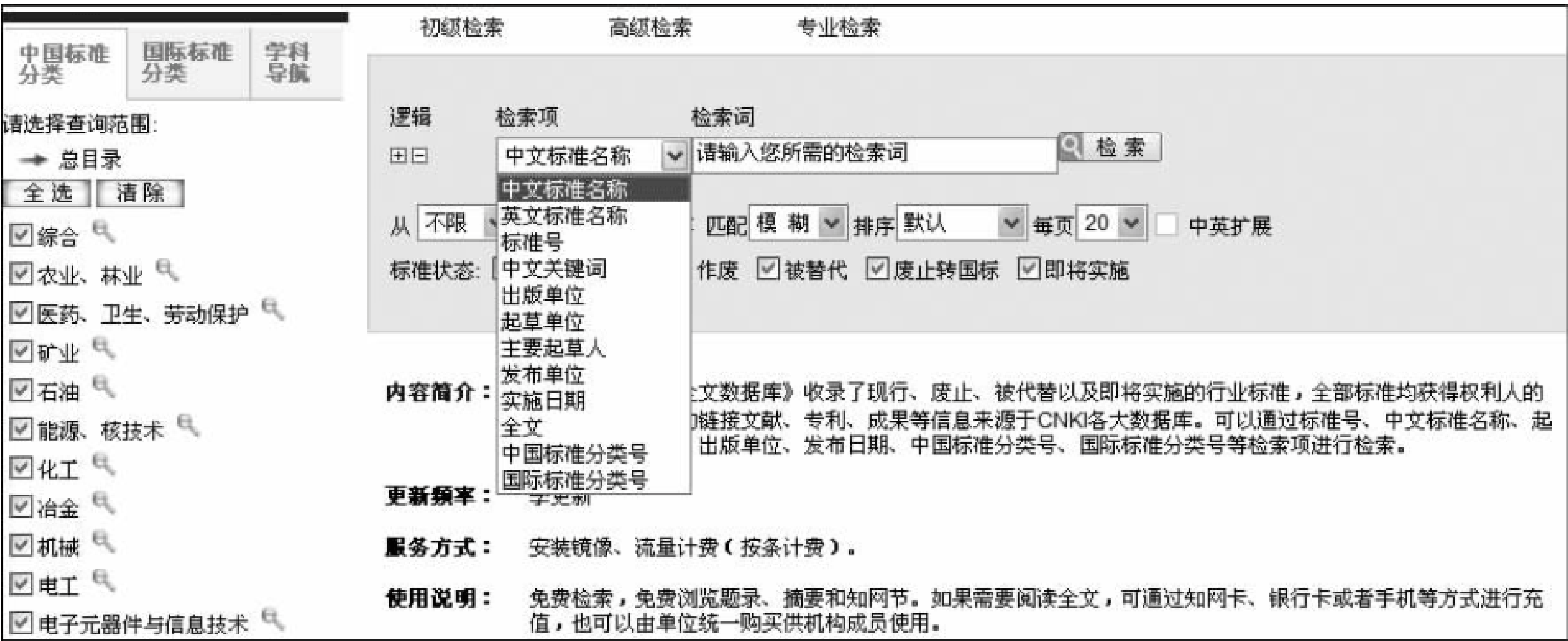


图 12-69 CNKI 标准的初级检索界面

版单位、实施日期等检索项进行逻辑组合。在检索复杂的标准时最多可以使用六个检索项进行高级逻辑组配。高级检索界面见图 12-70。

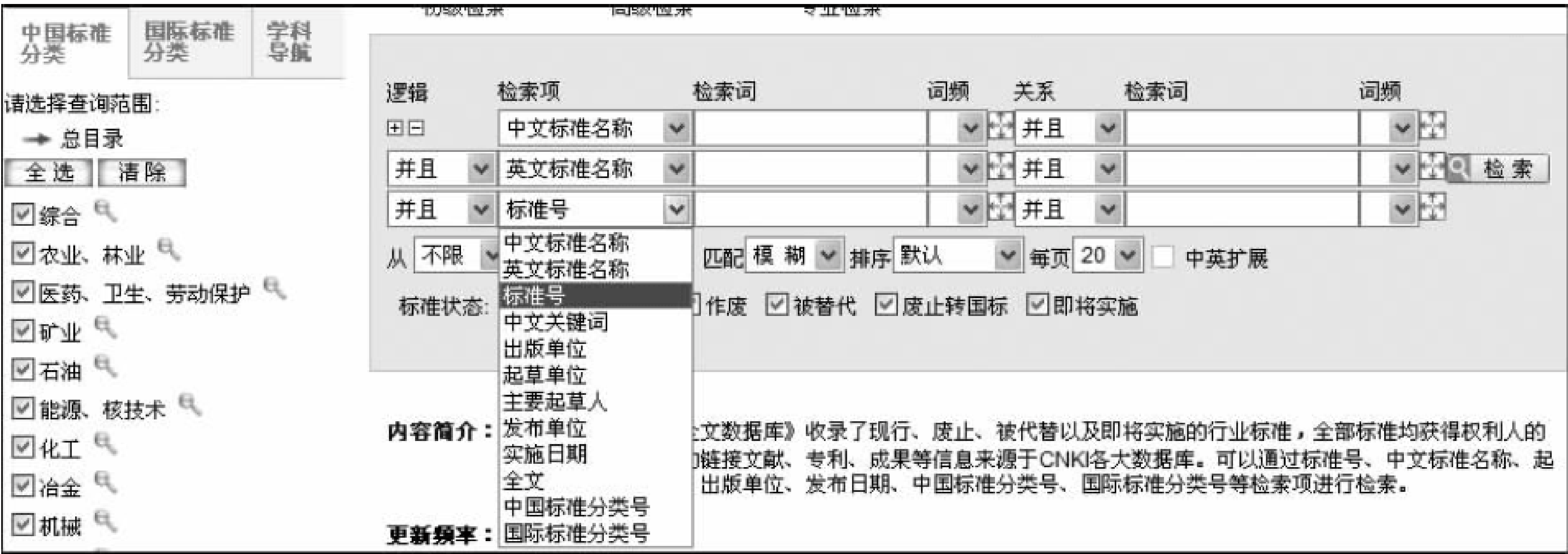


图 12-70 CNKI 标准的高级检索界面

(3) CNKI 标准的专业检索。用复杂逻辑表达式可以对中文标准名称、英文标准名称、标准号、机标关键词(中文关键词)、出版单位、起草单位、主要起草人、发布单位、实施日期、全文、中国标准分类号、国际标准分类号等内容进行逻辑表达,实现标准信息的精确检索。专业检索界面见图 12-71。

在使用方法方面:免费检索,免费浏览题录摘要和知网节,标准的全文下载需付费,请先注册(作为大学生,如果所在高校购买了标准数据库,查询时无须注册和付费)用户的个人账户,并通过知网卡、银行卡、神州行卡等方式给自己的账户充值。流量计费产品全文分为阅读版、打印版、阅读打印版三个版本,各版本使用方式及计费价格不同,请按提示下载。阅读版:只可阅读不可打印;打印版:只可打印三次不可阅读;阅读打印版:可阅

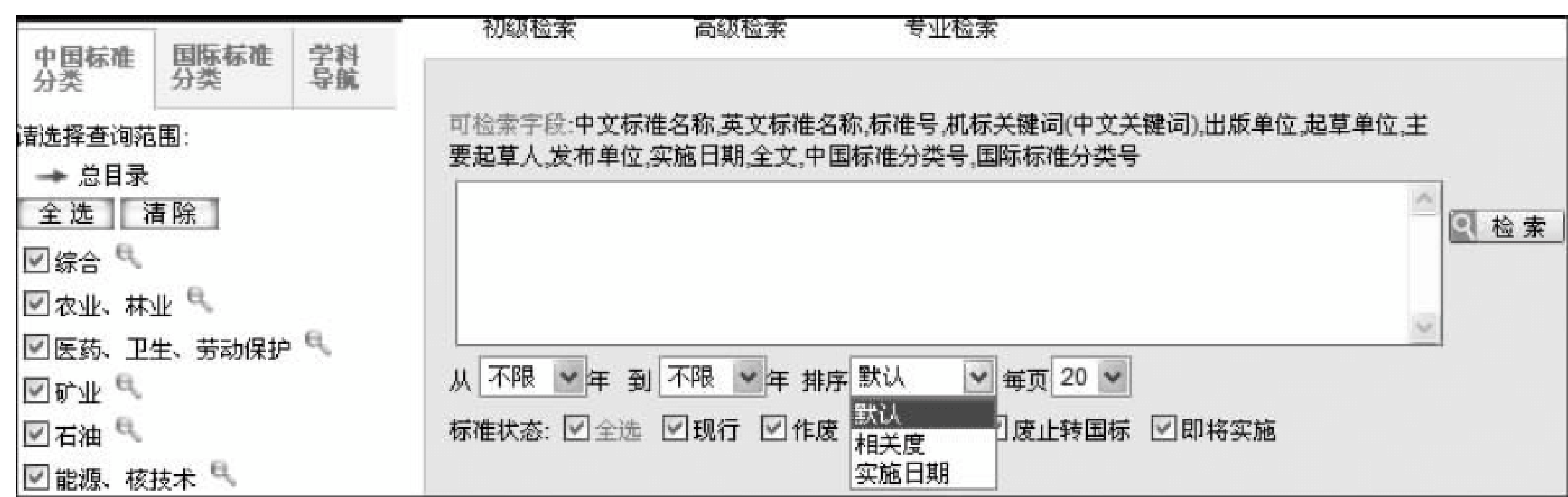


图 12-71　CNKI 标准的专业检索界面

读可打印三次。下载后只可在本机使用。第一次打开 PDF 全文时,请按提示下载安装 Adobe Acrobat 插件,插件安装成功后方可打开全文。

本 章 小 结

对于大学生或科技工作者而言,特种信息资源是指出版发行和获取途径都比较特殊的科技类信息资源,通常也指的是除了普通图书信息资源和期刊信息资源之外的特种科技信息资源。它们通常包括会议文献信息资源、科技报告信息资源、专利信息资源、学位论文信息资源、标准信息资源、科技档案信息资源、政府出版物信息资源等七大类。特种信息资源特色鲜明、内容广泛、数量庞大、学习与研究及其参考价值高,在整个信息资源与信息检索及其利用过程中起着非常重要的作用。特种信息资源的载体形式丰富,除了光盘与印刷型纸质载体外,目前大多数也以网络数据库的形式提供检索服务。

科技报告(Scientific& Technical Report)是指对科学、技术研究成果或研究进展的记录,也称研究报告或报告文献。科技报告的出现早于科技期刊,在科学交流制度化之前科技工作者们就已经生成各类科技报告了。但是,作为一种传递科技信息的特定类型的信息资源,其历史能追溯到 20 世纪初。当时,只是研究者或设计单位向经费资助机构提交关于研究或设计任务完成情况以及财务支出情况的报告,大量的研究成果以内部报告交流的形式出现。科技报告通常划分为:初期报告(Primary Report)或开题报告,是研究机构对研究项目的一个计划性报告;中期报告或过程报告,如研究过程中的现状报告、预备报告、中期报告、进展报告、非正式报告;结题报告或总结报告,即研究工作结束时的报告,如总结报告、综述报告、试验结果报告、竣工报告、公开报告等。

会议文献(Conference Literature)就是指在学术会议上宣读和交流的论文、报告及其

他有关资料，并且多数以会议录(Proceeding)的形式出现。世界上每年产生的会议论文约10多万篇，每年出现的各种会议录就达3000余种。主要的会议资源检索工具或检索平台包括《世界会议》、《会议论文索引》、《科技会议录索引》、中国学术会议文献数据库、中国重要会议论文全文数据库、中国学术会议在线等。

学位论文是高等院校和科研院所的本科生、研究生为获得学位资格(博士学位、硕士学位和学士学位)而撰写的学术性较强的毕业研究论文，英国称为Thesis，美国称为Dissertation。学位论文通常都是经过悉心指导，符合授予学位的要求，不少论文选题新颖，论述系统，见解独到，具有独创性，特别是博士论文，探讨一些前人没有论及过的新领域，并且提出具有独特、创新的见解。因此，学位论文是学者、专家及博士与硕士生智慧的结晶，是了解国内外科技研究发展的重要的信息媒介，是各国拥有自主知识产权的重要信息资源和知识宝藏，具有重大的开发利用价值。

专利文献是科学技术的宝库。它融技术、法律和经济信息于一体，是各单位各部门领导了解掌握国内外技术发展现状，进行技术预测和做出科学决策的依据，是科研人员和工程技术人员进行课题研究，解决技术难题不可缺少的工具；是发明人寻找技术资料，不断做出新的发明创造的源泉。专利文献(Patent Literature)是指记录有关发明创造信息的文献。广义包括专利申请书、专利说明书、专利公报、专利检索工具以及与专利有关的一切资料；狭义仅指各个国家或地区的专利局出版的专利说明书或发明说明书。

狭义的标准信息资源是指按规定程序制定，经公认权威机构或主管机关批准的一整套在特定领域内必须执行的规格、规则、技术要求等规范性文献资料，简称标准。标准是大学生获取的一种重要学习与参考资源类型。广义的标准指与标准化工作有关的一切信息资源，包括标准形成过程中的各种档案、宣传推广标准的手册及其出版物、揭示报道标准文献信息的目录、索引等。国外标准信息资源经常使用的名称有标准(standard)、规格(Specification)、公报(Bulletin)、建议(Recommendation)、法规(Code)、手册(Handbook)、规则(Rules Instruction)和工艺(Practice)等。标准的制定和类型按使用范围划分有国际标准、区域标准、国家标准、专业标准、地方标准、企业标准。

本章思考与练习题

1. 什么是特种信息资源？有哪几种基本类型？
2. 什么是科技报告？有哪些特征？
3. 科技报告有哪些类型？请举例说明。

4．国家科技成果网有哪些基本检索方式？举例说明其高级检索的基本应用方法。
5．万方中文科技报告数据库的基本检索方式有哪些？请举例说明。
6．举例说明国务院发展研究中心报告(国研报告)的检索方式。
7．举例说明中国商业报告数据库的高级检索与专业检索在检索方法方面的差异。
8．使用什么数据库可以对国外科技报告进行便捷检索？
9．什么是会议文献？有哪些类型和主要特点？
10．国外有哪些主要会议索引文献资源？
11．通过哪些数据库平台可以便捷检索国内会议文献资源？请举例说明。
12．什么是学位论文？如何检索国外主要学位论文全文信息？
13．有哪些主要数据库平台可以检索国内学位论文信息资源？
14．什么是专利文献？它有哪些基本类型？
15．专利文献检索有哪些主要作用？
16．有哪些主要国外专利信息检索数据库？
17．专利信息资源检索有哪些主要字段？请举例说明。
18．举例说明专利搜索引擎的主要检索应用。
19．专利高级检索与表格式检索的差异？请举例说明。
20．中国专利文献检索数据库平台有哪些？
21．举例说明如何应用高级检索功能检索中文专利文献信息。
22．什么是标准？标准文献有哪些类型？标准信息检索有何作用？
23．有哪些主要中外标准信息检索平台？
24．如何使用高级检索功能检索标准信息？

第 13 章　图书与学术期刊论文信息资源检索

图书是以传播知识为目的,用文字或其他信息符号记录于一定形式的材料之上的著作物;图书是人类社会实践的产物,是一种特定的不断发展着的知识传播工具。图书的基本构成要素有被传播的知识信息、有记录知识内容的文字或图像的信号、有存储与传播知识信息的物质载体、有图书的特定生成技术和工艺。图书的含义十分丰富,图书一般指书籍,由出版社出版的相对独立的出版物;有特定的书名和著(编)者名;每种书有不同的篇幅(印张)和不同的定价,并标有国际图书标准书号 ISBN。图书一般不做广告,但可以重印和修订再版。图书主要分为社会科学和自然科学两大类。本章所指的是其狭义概念即书籍,即大学生能够通过图书馆或网络查询并获取的纸质与数字化图书。

期刊,也称杂志。《辞海》中期刊的定义是:定期或不定期的连续出版物。每期版式基本相同,有固定名称。用卷、期或年、月顺序编号出版,有专业性和综合性两大类。期刊,由杂志社定期出版的连续出版物,如半月刊、月刊、双月刊和季刊等。刊物有固定的名称、固定的印张和固定的定价,并使用国际标准期刊号(连续出版物号)ISSN;可设有多个栏目,版式比较活泼,内容包罗万象,并可做广告。刊物出版后一般不重印,但可制作合订本。期刊内容一般比较杂,故又称杂志,期刊分专业性和综合性两大类。本书所指的期刊是对大学生的自主学习、协作学习、探究性学习有辅助作用的学术期刊。

图书与期刊的主要区别是:期刊使用的是 ISSN 即国际标准期刊号(International Standard Serial Number,ISSN),俗称连续出版物号;图书使用的是 ISBN,即国际标准图书号(International Stan-dard Book Number,ISBN)。

13.1　大型中文图书目录检索系统

13.1.1　中国国家图书馆联机公共目录查询系统

中国国家图书馆,是世界五大藏书过千万册的图书馆之一,中国国家图书馆分为总馆南馆、总馆北馆和古籍馆,馆藏书籍 3119 万册,其中古籍善本有 200 余万册。2008 年中国国家图书馆建筑面积为 28 万平方米,是亚洲规模最大的图书馆,居世界国家图书馆第

三位。读者查询书籍可以使用“中国国家图书馆联机公共目录查询系统”。

1. 读者的个性需求“查询参数设置”

①“每页显示记录数”设置，可选择 3、10、15 或 20 条；②“自动完整显示记录数”设置，可选择 0、5、10 或 15 条；③检索分馆选择，默认为全部，可以选择中文图书借阅区、北区图书借阅区、南区工具书借阅区、古籍馆中文图书借阅区等 35 个具体馆藏部门资料；④查询数据显示格式设置，包括详细格式、题名格式、简明格式、卡片格式等；⑤列表数据是否包含规范数据设置。读者的个性需求“查询参数设置”如图 13-1 所示。



图 13-1　中国国家图书馆读者查询参数个性设置

2. 多语种虚拟键盘

使用关键词可以查阅外文图书，在查阅外文图书时，可随时启用多语种键盘快速输入查询词，例如选择日文平假名、俄文、希腊文等。多语种虚拟键盘使用如图 13-2 所示。

3. 检索限制

检索限制就是限制一定的图书查询范围：图书资源的语种限制，可以指定查询的图书为中文、英文、俄文、日文、德文或法文；限制图书出版的起止时间范围，例如 2010—2016；限定资料类型为图书、期刊、电子文献等；资料馆藏位置限定，例如法律参考阅览室。检索限制查询界面见图 13-3。

4. 检索字段限制

检索字段限制是对图书的书名、著者、分类号、主题词、出版单位、索取号、ISBN 号等进行限定以提高检索精确度。图 13-4 是用著者字段以“谭浩强”为检索词所获得的检索结果。

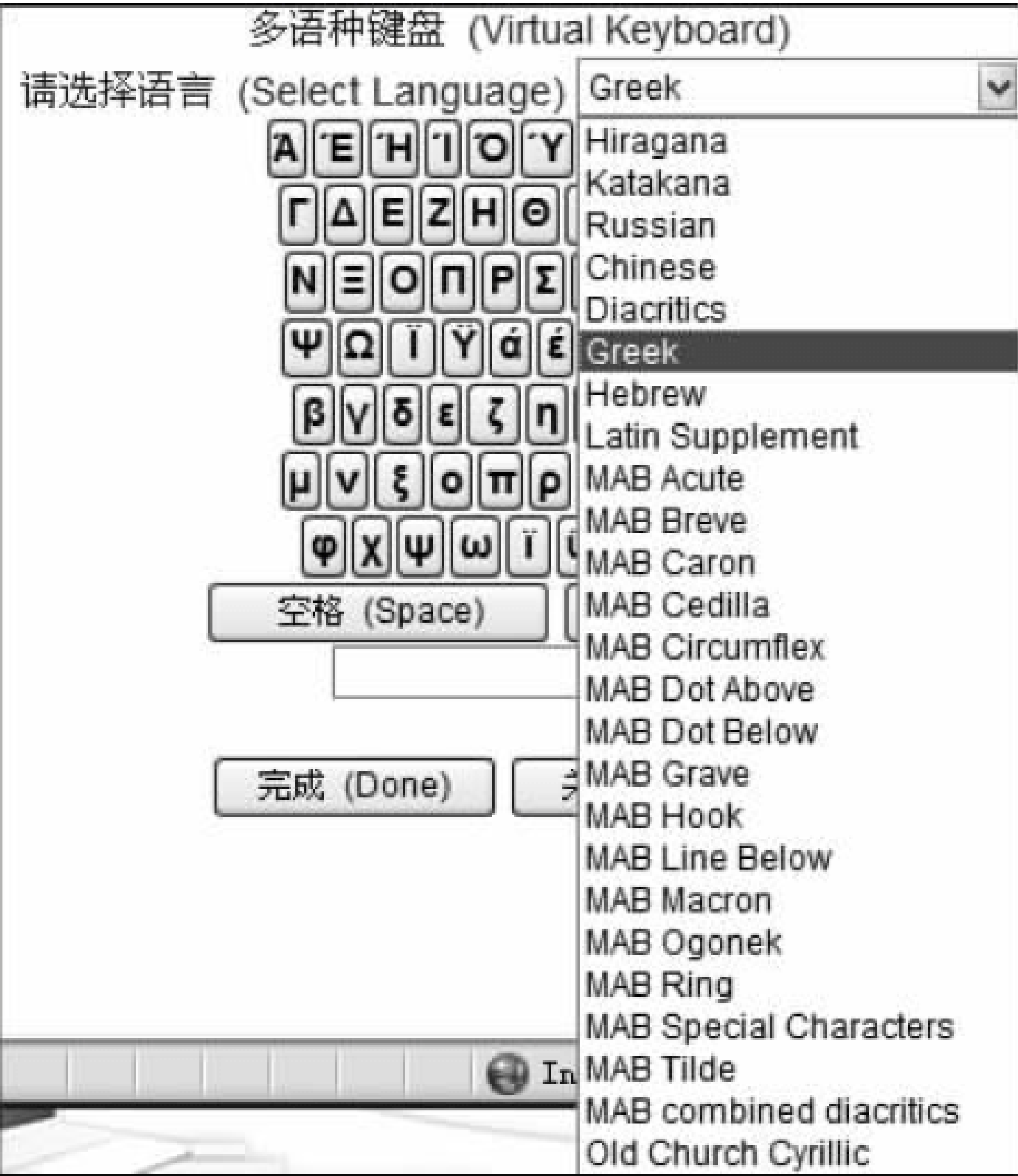


图 13-2　中国国家图书馆图书查询的多语种虚拟键盘

检索限制:
语言: 全部　开始年份:　结束年份:　yyyy (当不使用起/止时，使用？作截词)
资料类型: 图书　分馆　古籍馆普通古籍阅览室

图 13-3　中国国家图书馆检索限制查询界面

著者　谭浩强　中文文献库　搜 索
排序：年(降序)/著者　格式：封面视图　选中记录　整合集合　重新查询　分类浏览
记录 1 - 20 of 278 (最大显示记录 1000 条)　1 2 3 4 5 6 7 8 9 10 下一页>

1 麦肯锡说，未来20年大机遇 [专著] / 理查德.多布斯(Richard Dobbs)，詹姆斯
作者：(美)多布斯 (Dobbs, Richard) 著　出版社：广东人民出版社
年份：2016　格式：图书
馆藏复本: 4，已出借复本: 1

2 C++程序设计题解与上机指导 [专著] / 谭浩强编著
作者：谭浩强 (1934~) 编著　出版社：清华大学出版社
年份：2015　格式：图书
馆藏复本: 4，已出借复本: 1

3 C++程序设计 [专著] / 谭浩强编著
作者：谭浩强 (1934~) 编著　出版社：清华大学出版社
年份：2015　格式：图书
馆藏复本: 4，已出借复本: 1

图 13-4　检索字段检索实例

图 13-4 中可以运用“排序”控件对检索结果进行五种不同方式的结果排序：①著者/题名；②著者 /年(降序)；③年(降序) /著者；④题名/年(降序)；⑤年(降序)/题名。同时可以选择图书检索结果的不同输出格式，以适应读者的不同需求风格，例如“封面视图”、“简洁视图”、“详细视图”等。下面以 2015 年清华大学出版社的“C++ 程序设计［专著］/ 谭浩强编著”为例，说明几种不同的图书检索输出格式。见图 13-5～图 13-7。

C++程序设计 [专著] / 谭浩强编著 文献索取

作者：	谭浩强 (1934~) 编著	出版社：	清华大学出版社
年份：	2015	格式：	图书
			馆藏复本: 4，已出借复本: 1

图 13-5　图书检索结果的“封面视图”输出格式实例

谭浩强 (1934~) 编著	BK	C++程序设计 [专著] / 谭浩强编著	2015	中文图书借阅区(1/ 1) 中文基藏(1/ 0) 书刊保存本库(1/ 0) 北区中文图书区(1/ 0)

图 13-6　图书检索结果的“简洁视图”输出格式实例

ID 号	008009426
通用数据	20151008d2015 em y0chiy50 ea
题名与责任	C++程序设计 [专著] / 谭浩强编著
版本项	3版
出版项	北京 : 清华大学出版社, 2015
载体形态项	19,466页 ; 26cm
语言	chi
一般附注	中国高等院校计算机基础教育课程体系规划教材/谭浩强主编
内容提要	本书作者深入调查了大学的程序设计课程的现状和发展趋势，参阅了国内外数十种有关C++的教材，认真分析了读者在学习中的困难和认识规律，设计了读者易于学习的教材体系，于2004年出版了《C++程序设计》一书。此次"第3版"降低入门起点，不需要C语言的基础，从零起点介绍程序设计和C++。书中分为4篇：基本知识、基于过程的程序设计、基于对象的程序设计和面向对象的程序设计。
主题	C++语言 -- 程序设计 -- 高等学校 -- 教材
中图分类号	TP312C++
著者	谭浩强 (1934~) 编著

图 13-7　图书检索结果的“详细视图”输出格式实例

5. 高级检索

中国国家图书馆图书信息“高级检索”分为多字段检索、多库检索、组合检索、通用命令语言检索、(普通)浏览和分类浏览共六种。

(1) 图书多字段检索。它主要是对图书的主题、著者、题名起始于、出版年、词邻近关系

等进行多字段检索,实现比较精确查找图书资料的目的。多字段检索界面实例见图 13-8。

多字段检索	
主题	
著者	
题名起始于	(以题名的第一个字开始)
题名	
出版年	
出版者	
词邻近	○否 ⊙是
书目库	中文文献库
确 定	清 除 多语种键盘

图 13-8 图书的多字段检索界面实例

(2)图书多库检索。在检索时,可以对多个数据库同时展开检索,实现跨库检索的目的。在选择多个检索库时,也可以对检索资料类型范围、资料时间范围和物理馆藏范围进行检索限制。多库检索界面实例见图 13-9。

多库检索		
输入检索词或词组		
检索字段	其他题名	
词邻近?	○否 ⊙是	
	确 定 清 除	
选择数据库:		
☑中文及特藏数据库		
☐中文普通图书库	☐音像制品和电子资源(含中外文)	☐民语文献 全部
☐中文期刊	☐中文报纸	☐中文缩微文献
☐台港图书及海外出版的中文图书	☐普通古籍(含新线装)	☐善本古籍文献
☐学位论文	☐联合国资料	☐地方志、家谱文献
☐外文文献数据总库 **语种** ALL		
☐外文图书	☐外文善本	☐外文缩微文献
☐外文期刊	☐外文报纸(含台港外文报纸)	☐外文乐谱
☐外文地图	☐国际组织和外国政府出版物	

检索限制:			
开始年份:		结束年份:	yyyy (当不使用起/止时,使用 ? 作截词)
资料类型:	全部	分馆	全部

图 13-9 图书的多库检索界面实例

(3) 图书组合检索。组合检索就是对图书的多个检索字段进行逻辑组配(例如逻辑与)检索。在组合检索时,可以对查询资料的最多三个字段进行逻辑组合操作,同时进一步对“检索词邻近否”和“检索限制”条件进行控制。组合检索界面实例见图 13-10。

图 13-10　图书的组合检索界面实例

(4) 图书通用命令语言检索。通用命令语言检索也通常称为“专业检索”,中国国家图书馆图书的通用命令语言检索的主要方法有以下几种。

① 主要检索命令。例如“WRD=(计算机 OR 电脑)AND 软件”,将检索出包含计算机或电脑且包含软件的信息记录。WRD——任意字段,WTI——题名字段,WAU——作者字段,WSU——主题字段,WPU——出版者字段,WYR——出版年字段。

② 词邻近否的含义。词邻近选择为“是”,表示检索词或短语完整地出现在检索字段中。词邻近选择为“否”,表示检索词可以分开位于所检索的字段中。没有选择“是”或“否”,系统将以上次检索的值为默认选择进行检索。

③ 检索词中的标点。检索词中的标点符号应当去掉,如. 号等。例如 visual basic 6. 0 中的点,应在检索时去掉,输入为 60 即可。

④ 外文图书的作者。外文文献的作者姓名输入顺序为:姓在前,名在后。如 Bill Gates 的正确输入为“Gates Bill”,而不是“Bill Gates”,“Bill · Gates”,“Bill, Gates”,“Gates, Bill”等。

⑤ 逻辑运算的默认。and(与)为检索词之间的默认逻辑运算。如果需要使用其他逻

辑操作,可以选择通用命令语言方式。

⑥ 通配符? 和 * 的应用。? 或 * 可用于单词的开始或结尾,代替单词的其他部分。?ology 检索到 anthropology,archaeology,psychology 等。Chloro? 检索到以 Chloro 开头的单词。? 查找不同的拼写方式。如 alumi? m 可以匹配美式拼写 aluminum 和英式拼写 aluminium。? 不能同时用于单词的开始和结尾,如 ? dva? 视为非法。? 或 * 作为占位符,可以代替任意多个字符。如 ps? ic,检索到以 ps 开头、以 ic 结尾的所有单词。

⑦ 通配符%和! 的应用。%与一个数字连用,表示出现在两个检索词之间的单词个数小于该参数,检索词出现的顺序不固定。如 england %3 ballads 检索到:Ballads of England,England and Her Ballads,and Ballads of Merry Old England 等。! 与一个数字连用,表示两个检索词之间固定出现若干个单词,且检索词出现的顺序与输入顺序相同。如 ballads ! 3 england 可以检索到 Ballads of England,Ballads of Merry Olde England。但不会出现 England and Her Ballads。使用%和! 时,"词邻近"必须选择"是"。

通用命令语言检索界面实例见图 13-11。

图 13-11 图书的通用命令语言检索界面实例

(5)(普通)浏览检索。依据检索词的中文或西文顺序索引特征,对检索结果的列表进行浏览查询。浏览查询时,可以设定检索词为正题名、其他题名、主题词、著者等属性及检索词所属的范围为中文文献库或西文文献库。

(6)分类浏览检索。依据《中国图书馆图书分类法》对文献信息资源的主题分类目录,在分类目录的多级子目录中逐级浏览查询,来获得需要的查询对象。例如,逐级浏览 T 工业技术—TU 建筑科学—TU5 建筑材料—TU52 非金属材料—TU523 建筑陶瓷及制品,可获得相应图书文献资料。

13.1.2　CALIS 联合目录公共检索系统

中国高等教育文献保障系统(China Academic Library & Information System, CALIS),是经国务院批准的我国高等教育“211 工程”“九五”“十五”总体规划中三个公共服务体系之一。CALIS 把国家投资、现代图书馆理念、先进技术手段、高校丰富的文献资源和人力资源整合起来,是一个以中国高等教育数字图书馆为核心的教育文献联合保障体系,实现信息资源共建与共享。

(1) 基本检索方法。CALIS 联合目录公共检索系统(以下简称 OPAC)采用 Web 方式提供查询与浏览。

① 多库分类检索:OPAC 中的数据,按照语种划分可分为中文、西文、日文、俄文四个数据库;按照文献类型划分,可分为图书、连续出版物、古籍。

② 排序功能:默认的排序优先次序是题名、相关度。

③ 检索历史:保留用户发出的最后 10 个检索请求,用户关闭浏览器后,检索历史将清空。

④ 多种显示格式:检索结果分为多种格式显示,包括详细文本格式、MARC 显示格式。前一种格式对所有用户免费开放,MARC 显示格式只对 CALIS 联合目录成员馆开放,查看或下载 MARC 记录,均按照 CALIS 联合目录下载费用标准收取。

⑤ 多种格式输出:对所有用户提供记录引文格式、简单文本格式、详细文本格式的输出,此外,对 CALIS 联合目录成员馆还提供 ISO2709、MARCXML、CALIS bookXML、MARC 列表的输出。提供 E-mail 与直接下载到本地两种输出方式。输出字符集提供常用的“GBK”、“UTF-8”、“UCS2”、“MARC8”四种,用户可根据自己的需要进行选择。

⑥ 浏览功能:对古籍数据提供四库分类的树形列表浏览。

⑦ 收藏夹功能:对有权限的用户提供保存用户的检索式与记录列表、标注书签、添加和维护用户评论的功能,目前这些功能不对普通用户开放。

⑧ 馆际互借:OPAC 系统提供用户直接发送请求到本馆的馆际互借网关,用户无须填写书目信息。

(2) 简单检索。默认为“全面检索”,也可以选择题名、责任者、主题、分类号、ISBN 号等检索项。简单检索界面见图 13-12。

(3) 高级检索。高级检索就是对多个检索项(例如题名、作者、出版者等)进行与、或、非的布尔逻辑表达且进一步组配检索项的“包含”、“前方一致”与“精确匹配”关系,实现检索的高查准率。高级检索界面见图 13-13。

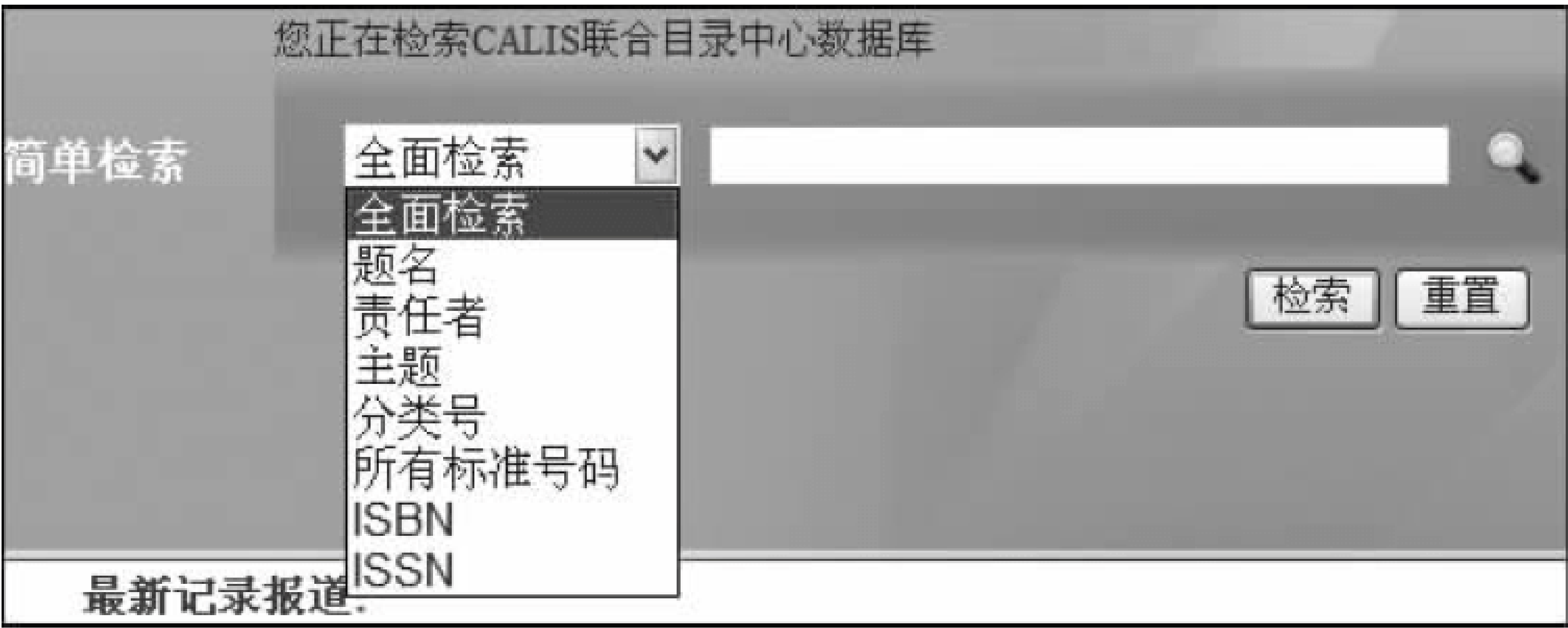


图 13-12　CALIS 的简单检索界面



图 13-13　CALIS 的高级检索界面

CALIS 高级检索的一般方法包括：①选择检索点，输入检索词，选择限定信息，单击“检索”按钮或直接按 Enter 键；②默认的检索匹配方式为前方一致，也可以在复选框中选择：精确匹配或包含；③最多可输入三项检索词，默认逻辑运算方式为“与”，也可以在复选框中选择“或”、“非”；④选择分类号检索点，可以单击“中图分类号表”按钮浏览，选中的分类号将自动填写到检索词输入框中；⑤限制性检索的文献类型可选择普通图书、连续出版物、中文古籍，默认为全部类型；⑥限制性检索的内容特征可选择：统计资料、字典词典、百科全书，默认为全部；⑦可通过输入出版时间对检索结果进行限定。例如，选择“介于之间”并输入“1998—2000”，即检索 1998 年至 2000 年出版的文献；⑧检索词与限制性检索之间为“与”的关系。

13.1.3　北京大学图书馆公共查询系统

北京大学图书馆(Peking University Library)是中国最早的现代新型图书馆之一，是

我国最大的综合性高等教育图书馆，已发展成为资源丰富、现代化、综合性、开放式的研究型图书馆。截至 2015 年年底，北京大学图书馆由总馆、医学馆、38 个分馆、储存馆组成；总、分馆文献资源累积量约 1100 万册（件），其中纸质藏书 800 余万册，以及大量引进和自建的国内外数字资源，包括各类数据库、电子期刊、电子图书和多媒体资源约 300 余万册（件）。北京大学图书馆公共查询系统分为基本检索和高级检索两种途径，资源类型分为图书和期刊两大类。

（1）基本检索。图书检索的时间范围包括全部时间图书、最近三天新书、最近一周新书和最近一月新书；检索模式默认为任意匹配，也可以选择完全匹配、前方一致和后方一致。基本检索界面见图 13-14。



图 13-14　北京大学图书馆公共查询系统的基本检索界面

（2）高级检索。可以对图书的六个主要检索项（ISBN、正题名、出版社、主题词、责任者和出版年）进行逻辑组合检索。高级检索界面见图 13-15。

13.1.4　清华大学图书馆馆藏目录检索系统

清华大学图书馆是我国大型高校图书馆之一，馆藏资源十分丰富。截至 2015 年年底，清华大学图书馆（含专业图书馆及院系资料室）的实体馆藏总量约 491.2 万册（件），形成了以自然科学和工程技术科学文献为主体，兼有人文、社会科学及管理科学文献等多种类型、多种载体的综合性馆藏体系。除中外文印刷型图书外，读者可使用的文献资源还包括古籍线装书 22 万多册、期刊合订本约 57.4 万册、校馆统筹年订购印刷型中外文报刊



图 13-15　北京大学图书馆公共查询系统的高级检索界面

2394 种、学校博硕士论文 14.3 万余篇、缩微资料 2.8 万种、各类数据库 551 个、全文电子期刊 69 737 种、电子图书 840.3 万册、电子版学位论文 353.3 万篇。通过清华大学图书馆馆藏目录检索系统可查询图书馆收藏的中西文图书、日文图书、俄文图书、中西文期刊和 1994 年以后入藏的日文期刊、多媒体资源、大部分外文电子期刊、学位论文和中外文电子图书，以及六个专业图书馆和部分院系资料室的馆藏。古籍通过馆藏古籍目录查询，其余馆藏文献通过卡片目录查询。通用检索界面见图 13-16。

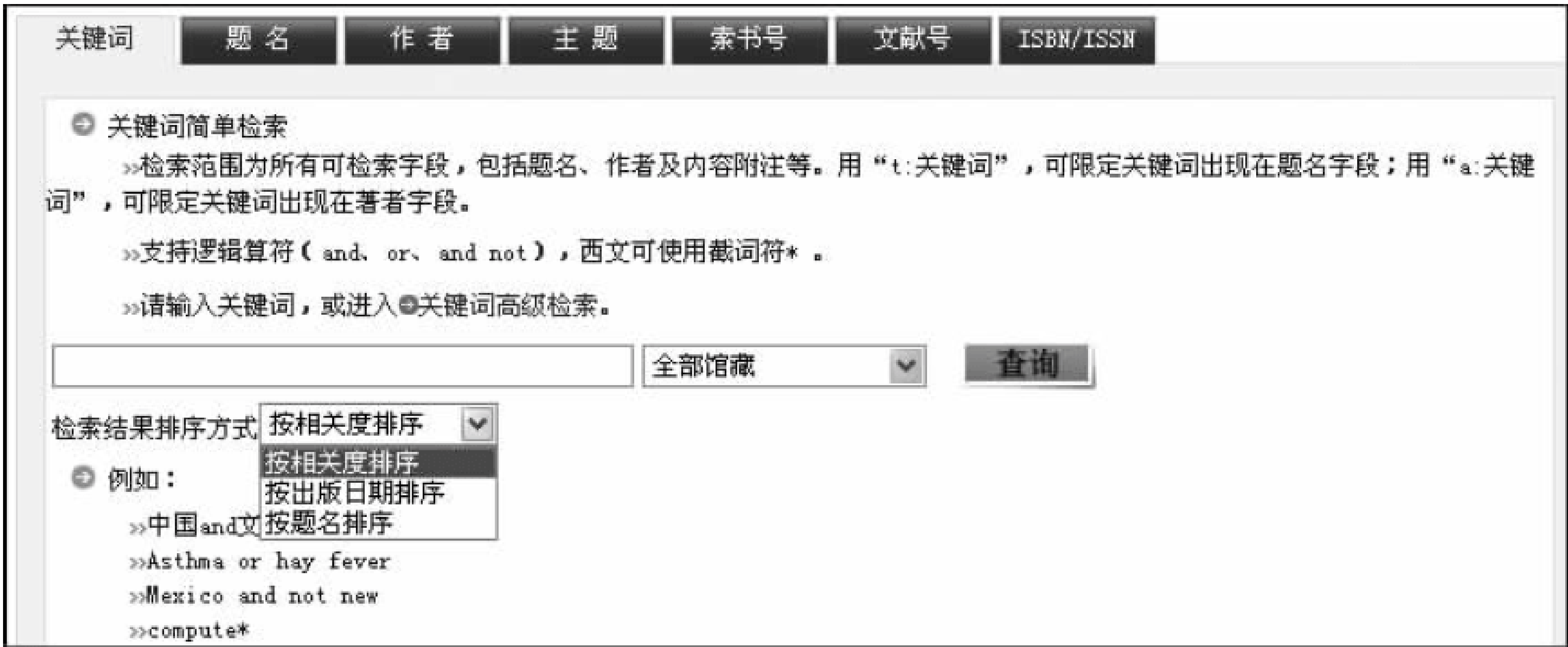


图 13-16　清华大学图书馆馆藏目录检索系统的通用检索界面

1. 关键词高级检索

支持逻辑算符(and、or、and not)，西文可使用截词符 *，可以最多对四个检索项进行逻辑组合(逻辑与、或、非)检索，同时可选择馆藏范围(例如文科馆、法律馆、总馆等)和资源类型(纸质图书、期刊、工具书、电子资源等)，也可以控制检索结果的输出语种(例如中文、法文、德文、英文等)。关键词高级检索界面见图 13-17。



图 13-17　关键词高级检索界面

2. 关键词简单检索

检索范围为所有可检索字段,包括题名、作者及内容附注等。用“t:关键词”,可限定关键词出现在题名字段;用“a:关键词”,可限定关键词出现在著者字段。关键词简单检索界面见图 13-18。



图 13-18　关键词简单检索界面

3. 其他检索方式

其他检索方式包括题名、作者、主题词、索书号、文献号和 ISBN 等。

(1) 题名检索。系统默认检索方式为前方一致,请输入完整题名或题名起始部分,题名包括图书名、期刊刊名、丛书名等出版物名称。如果想使检索词出现在题名的非起始部分,请通过“关键词”途径,用“t:关键词”语法实现。例如,完整图书名: Gone with the wind;完整丛书名: 中国房地产研究丛书;图书名的起始部分: Gone with;丛书名的起始部分: 中国房地产。

(2) 作者检索。作者检索范围为个人作者、团体作者和会议名称等。对于个人作者

请先输入姓，对于团体作者或会议名称请输入名称的缩写或起始部分。例如，个人作者全名(姓在前)：Smith,John;个人作者名起始部分：Smith,J;个人作者名起始部分：Smith;团体作者全名：清华大学;团体作者名起始部分：清华;完整会议名称：Institute of Electrical and Electronics Engineers;会议名称缩写：IEEE。

(3) 主题词检索。检索范围为主题词字段;主题词是用来揭示资料内容特征的词或词组;主题词来源于词表，中文出版物的主题词选自汉语主题词表，西文出版物的主题词选自美国国会图书馆主题词表。例如，Sports medicine、Sports、计算机、计算机——软件。

(4) 索书号检索。索书号由分类号和区分号构成，分类号与区分号用空格隔开。要查一本书可输入完整的索书号，要查一类书可输入分类号。例如，H316 FA51、TP316 25、H316。

(5) 文献号检索。OPAC 中的文献号包括文献标识的多种代号(码)，如文献的国家书目号、版权登记号、政府出版物号、标准技术报告号、CODEN 代码、统一书刊号、标准号、中文图书订购号、中文期刊的 CN 号等。请输入完整的文献号或文献号的起始部分。例如，730B0001、CN 11-1018。

(6) 国际标准书号检索。国际标准号码检索包括 ISBN、ISSN、ISRC 等。ISBN 为国际标准书号，ISSN 为国际连续出版物号，ISRC 为音像制品国际标准编码。例如，10 位 ISBN：7-5354-3028-7;10 位 ISBN：7535430287;13 位 ISBN：978-7-5063-4321-3;13 位 ISBN：9787506343213。

13.2 典型中文数字图书检索——超星数字图书馆

“超星数字图书馆”为目前最大的中文在线数字图书馆，提供大量的电子图书全文资源供阅读，其中包括文学、经济、计算机等五十余大类，总数 120 多万种电子图书，500 万篇论文，全文总量 13 亿余页，超 16 万集的学术视频。超星数字图书馆成立于 1993 年，是国内专业的数字图书馆解决方案提供商和数字图书资源供应商。超星数字图书馆，是国家“863”计划中国家数字图书馆示范工程项目，2000 年 1 月在互联网上正式开通，由北京世纪超星信息技术发展有限责任公司投资兴建。

1. 超星中文电子图书

高校图书馆大多购买了超星中文电子图书，一般分为学校镜像和远程访问两种形式，数字化近 300 家图书馆馆藏的近 120 万种全文电子书。

(1)超星阅读器安装。超星阅读器是超星数字化资源的专用阅读器，在手机端或 PC

端首先成功安装后，才能阅读超星全文电子资源。下载安装与提示见图 13-19。



图 13-19　超星阅读器 SSreader5.4 下载安装与提示

成功安装后，作为大学生因为不同的使用环境（例如校园网或非校园网环境）差异，提示界面略有差异，图 13-20 是成功安装后在桂林电子科技大学校园网中使用的界面。



图 13-20　超星阅读器成功安装后的初始界面

（2）图书分类检索。直接在超星阅读器 SSreader 左侧的一级分类目录中直接打开二级目录后查询。例如依据“图书分类”→“工业技术”→“自动化与计算机技术”的顺序，可以直接查询到马化腾著的《互联网＋国家战略行动路线图》一书。阅读全文的方式有两种：一是网页阅读，二是阅读器阅读。作为高校大学生用户，一般推荐阅读器阅读方式。见图 13-21。

在超星电子书“在线阅读”模式下，左侧为图书目录，读者可以通过左侧目录直接跳转查阅图书原文内容，也可以逐页阅读内容。该模式提供了图书内容的放大、缩小、文字摘录、打印、下载，同时提供三种全文电子图书阅读模式：带目录阅读、双页阅读和全屏连页阅读。见图 13-22。

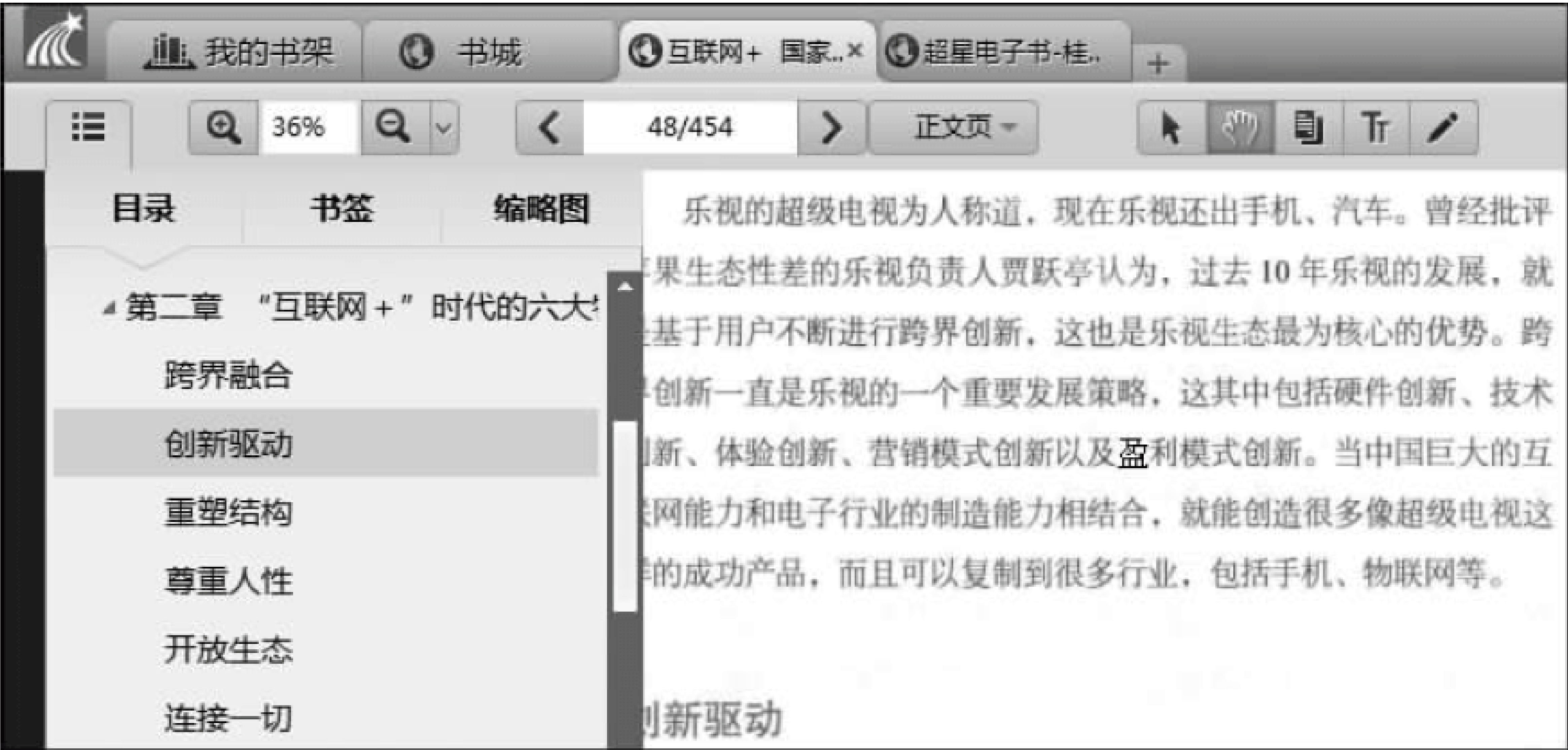


图 13-21 超星电子书“阅读器阅读模式”实例



图 13-22 超星电子书“在线阅读模式”实例

（3）搜索“我的书架”。读者可以将需要阅读的图书存放在“我的书架”模块中，便于直接从我的书架中查阅或直接搜索曾经阅读且需要继续完整或详细阅读的图书。见图 13-23。

（4）简单检索。直接使用书名、作者、目录和全文检索项进行检索，同时可以依据“图书出版日期”和“书名”进行检索结果排序。例如，使用“律师”作为书名进行检索，结果如图 13-24 所示。

（5）高级检索。同时对书名、作者、主题词、出版时间段、主题分类、分类号、搜索结果显示条数进行逻辑组配检索，以提高对电子图书的检索精度。见图 13-25。

2. 超星读秀中文学术搜索

“超星读秀中文学术搜索”是全球最大的中文图书搜索及参考咨询文献传递系统，目



图 13-23　超星电子书“我的书架”查阅实例



图 13-24　超星电子书一般检索实例



图 13-25　超星电子书高级检索实例

前收录310多万册图书数据、200多万种原文共8亿多页文献资料，提供全文检索、图书搜索及多种搜索功能，目的是让读者“找到得到”和“集天下之书为一书”。

(1) 图书普通检索。在搜索框直接输入关键词，关键词可定位到全部字段、书名、作者或主题词中，然后单击“中文搜索”按钮，将为用户在海量的图书数据资源中进行查找。如果希望获得外文资源，可单击“外文搜索”按钮。见图13-26。



图13-26　超星读秀一般检索视图

(2) 图书高级检索。在检索框输入图书的任一或多个检索项(例如ISBN、主题词、说明、出版时间等)进行逻辑组配，然后单击“高级搜索”按钮，更准确地定位到所需要的图书。见图13-27。



图13-27　超星读秀高级检索视图实例

（3）图书专业检索。专业标识符的使用含义：T＝书名，A＝作者，K＝关键词，S＝摘要，Y＝年，BKs＝丛书名，BKc＝目录。检索规则如下（以下符号均为半角符号）。

① 逻辑符号：* 代表并且，|代表或者，－代表不包含。

② 其他符号：() 括号内的逻辑优先运算，＝后面为字段所包含的值，＞代表大于，＜代表小于，＞＝代表大于等于，＜＝代表小于等于。

③ 大于小于符号仅适用于年代 Y，如果只有单边范围，字段名称必须写前边，如 Y＜2013，不允许写出 2013＞Y ；年代不允许单独检索。例如，题名或关键词中含有“图书馆”，且出版年范围是 2013 年至 2016 年（含边界），则专业检索表达式为：（T＝图书馆|K＝图书馆）*（2000＜＝Y＜＝2013），实例如图 13-28 所示。



图 13-28　超星读秀专业检索实例

3. 超星发现

超星发现以近十亿海量元数据为基础，利用数据仓储、资源整合、知识挖掘、数据分析、文献计量学模型等相关技术，较好地解决了复杂异构数据库群的集成整合，完成高效、精准、统一的学术资源搜索，进而通过分面聚类、引文分析、知识关联分析等实现高价值学术文献发现、纵横结合的深度知识挖掘、可视化的全方位知识关联，能够为大学生的探究性与研究型学习提供专业搜索服务。

（1）超星发现一般检索。直接用关键词、作者等单一检索项进行检索。图 13-29 是以“杂交水稻 袁隆平”为检索词的一般检索结果，包含了丰富的发现与分析数据（例如被引频次、研究趋势图等）。

（2）超星发现高级检索。不仅可以选择待检索资源的语种与文献类型，还可以通过“＋”和“－”符号来“按需调节”最大检索项，以实现精确检索。超星发现高级检索视图图 13-30。

（3）超星发现专业检索。包括以下几方面内容。

① 专业检索的通用字段标识符：T＝题名（书名、题名），A＝作者（责任者），K＝关键词，S＝文摘（摘要、视频简介），O＝作者单位（作者单位、学位授予单位、专利申请人），Su



图 13-29　超星发现一般检索实例图

＝主题，Z＝全部字段，Y＝年（出版发行年、学位年度、会议召开年、专利申请年、标准发布年）。

② 专业检索的文献类型标识符：BK＝图书，JN＝期刊，DT＝学位，CP＝会议，PT＝专利，ST＝标准，VI＝视频，NP＝报纸，TR＝科技成果。

③ 非通用字段标识符（需要加上文献标识才能检索）。图书：BKs＝丛书名；期刊：JNj＝刊名；学位：F＝指导老师，DTn＝学位，Tf＝英文题名，DTa＝英文文摘；会议：CPn＝会议名称；报纸：NPn＝报纸名称；专利：PTt＝专利类型；标准：STd＝起草单位。

④ 检索基本方法应用。包括以下几方面内容。

运算符号：* 代表并且，| 代表或者，－代表不包含，""代表精确匹配，"代表模糊匹配。

逻辑关系符：AND（与）、OR（或）、NOT（非）用于字段之间的逻辑关系，前后要空一

图 13-30　超星发现高级检索视图

个字节。

运算符及逻辑符的优先级相同，若要改变组合的顺序，请使用英文半角括号“()”括起；如：检索期刊题名包含图书馆或教育，且作者是王伟，出版年范围 2000 年至 2013 年(含边界)：JN(T=图书馆|教育 AND A=王伟) AND (2000<Y<2013)。

外文数据字段的值需要加模糊匹配符号" 或者精确匹配符号""，如 K="cryptography"|'cipher code'|"Multimedia security"(注：所有符号和英文字母，都必须使用英文半角字符)。

超星发现专业检索实例见图 13-31。

(4) 超星发现系统的核心搜索价值。激发创新灵感，洞察全局以发现科学研究价值，让巨人的肩膀成为知识价值再生的基石。

① 多维分面聚类。超星发现依托高厚度的元数据资源，通过采用分面分析法，可将



图 13-31 超星发现专业检索实例

搜索结果按各类文献的时间维度、文献类型维度、主题维度、学科维度、作者维度、机构维度、权威工具收录维度以及全文来源维度等进行任意维度的聚类。用户可根据实际需要进行任意维度的组配检索、自由扩检和缩检,从而实现文献资源发现的精炼聚类和精准化搜索,将最重要、最核心、最有价值的资源按相关度、被引频次、时间、影响因子等方式进行结果呈现。

② 智能辅助检索。超星发现提供强大的智能辅助搜索功能,借助内置规范知识库与用户的历史检索发现行为习惯,自动判别并切换到与用户近期行为最贴切的领域和关注热点,同步显示与用户检索主题相应的解释,帮助实时把握所检索主题的内涵,并优先按用户筛选文献的喜好显示检索结果,提高发现精准度和检准率。

③ 立体引文分析。超星发现可实现图书与图书之间、期刊与期刊之间、图书与期刊之间,以及其他各类文献之间的相互参考、相互引证关系分析。借助超星发现的文献引用频率分析研究,可有效测定与评价某一文献、某一学科、某一作者乃至某一机构的学术影响力。借助超星发现的文献间相互引证逻辑关系,可分析获得某一学术思想的历史渊源、传承脉络以及演变规律。

④ 探究学术源流。探究学术源流可以把文献资源的研究单位从单一的文献深化到文献中存在的知识关联中。通过学术源流可以按照知识概念形成知识相关链,这些关联就是知识关联的基础。超星发现能够按照知识概念给出知识关联图谱,通过单向或双向线性知识关联构成的链状、网状结构,形成主题、学科、作者、机构、地区等关联图,从而反映出学术思想之间的相互影响和源流。

⑤ 揭示知识关联。超星发现集知识挖掘、知识关联分析与可视化技术于一体,能够将发现数据及分析结果以表格、图形等方式直观展示出来。知识关联是我们从事知识活动和知识管理的基础,知识管理的目的是为科学组织和有效利用知识,而知识关联是科学

组织和有效利用知识的基本出发点和理论依据。因此,可以说知识管理的本质是知识关系的管理,通过知识关联为研究者从宏观角度直观地把握海量数据之间的规律和整体面貌,直观揭示人与人、人与机构、人与知识,以及知识与知识之间的关联,从而反映出不同学者、不同机构对某一领域的研究强度与贡献,反映出某一领域关联知识的相互交叉支持强度,为进一步追踪、拓展和创新该领域的研究提供思路。

⑥ 揭示学术趋势。超星发现具备对搜索结果进行年代分布规律分析的功能,可揭示出任一主题学术研究的时序变化趋势图,进而帮助研究者在大时间尺度和全面数据分析的高度洞察该领域研究的起点、成长、起伏与兴衰,从整体把握事物发展的完整过程和走向。无论是在上升或下滑趋势曲线中,当曲线在某一阶段处于上升或者处于波峰阶段时,即是在该时间段内学术研究兴盛的时段;当曲线在某一阶段处于下滑或者处于波谷阶段时,即是在该时间段内学术研究低迷的时段,同时也具有学术趋势发展的预判分析,为预测该学术未来发展的趋势提供帮助。

13.3　典型中文学术期刊论文检索

学术期刊(academic journal)是一种经过同行评审的学术性刊物,在学术期刊上发表的文章通常涉及特定的学科。学术期刊展示了某些研究领域的研究成果,并起到了公示的作用,其内容主要以原创研究、综述文章、书评等形式的学术文章为主。学术期刊论文也是大学生进行自主学习、探究与发现学习所不可或缺的重要参考资料,在学习过程中的课程小论文、实验报告、课程设计、实习报告、学术成果发表、创新与实践项目申报及其毕业论文撰写,都需要查阅专业学术期刊论文资料。

作为大学生,需要了解和把握自身专业领域的学术期刊,尤其是专业性的核心期刊。《中文核心期刊要目总览》(简称北大核心)由中国知网、中国学术期刊网和北京大学图书馆期刊工作研究会联合发布。中文核心期刊目录是学术界对某类期刊的定义,是一种期刊等级划分类型,它的对象是中文学术期刊,是根据期刊影响因子等诸多因素所划分的期刊。中文核心期刊是北京大学图书馆联合众多学术界权威专家鉴定,目前受到了学术界的广泛认同。从影响力来讲,其等级属同类划分中较权威的一种,是除南大核心、中国科学引文数据库以外学术影响力最权威的一种。按照惯例,北大核心期刊每四年由北大图书馆评定一次,并出版《北大核心期刊目录要览》一书。

国内核心学术期刊评选体系有:北京大学图书馆“中文核心期刊”、南京大学“中文社会科学引文索引(CSSCI)来源期刊”、中国科学技术信息研究所“中国科技论文统计源期

刊”(又称“中国科技核心期刊”)、中国社会科学院文献信息中心“中国人文社会科学核心期刊”、中国科学院文献情报中心“中国科学引文数据库(CSCD)来源期刊”、中国人文社会科学学报学会“中国人文社科学报核心期刊”以及万方数据股份有限公司建设的“中国核心期刊遴选数据库”。

13.3.1 CNKI 中国学术期刊网检索

国家知识基础设施(National Knowledge Infrastructure,NKI)的概念由世界银行于1998年提出。CNKI工程是以实现全社会知识资源传播共享与增值利用为目标的大型信息化建设项目,由清华大学、清华同方发起,始建于1999年6月。《中国学术期刊(网络版)》(国内统一连续出版物号CN11—6037/Z)是世界上最大的连续动态更新的中国学术期刊全文数据库,是“十一五”国家重大网络出版工程的子项目,是《国家“十一五”时期文化发展规划纲要》中国家“知识资源数据库”出版工程的重要组成部分。

CNKI中国学术期刊网的内容以学术、技术、政策指导、高等科普及教育类期刊为主,内容覆盖自然科学、工程技术、农业、哲学、医学、人文社会科学等各个领域。收录国内学术期刊8192种,全文文献总量46 759 660篇。学术论文数据库产品分为十大专辑:基础科学、工程科技Ⅰ、工程科技Ⅱ、农业科技、医药卫生科技、哲学与人文科学、社会科学Ⅰ、社会科学Ⅱ、信息科技、经济与管理科学。十大专辑下分为168个专题。数据库收录的论文为自1915年至今出版的期刊,部分期刊回溯至创刊。

(1) CNKI中国学术期刊网分类检索。依据检索界面左侧的分类导航目录逐级分类查找,可以获得子类中的学术论文资料。图13-32是依据“分类目录”→“信息科技”→“互联网技术”→“网络安全”的分类层次所获得的检索结果实例。

信息科技
无线电电子学
电信技术
计算机硬件技术
计算机软件及计算机应用
互联网技术
计算机网络理论
网络结构与设计
通信协议
通信设备与线路
网络管理与运行
网络安全
网络应用程序
各种网络

(0) 清除 导出/参考文献 分析/阅读 找到75,819条结果 浏览1/300 下一页

	篇名	作者	刊名	年/期	被引	下载	预览	分享
1	基于OpenFlow的SDN技术研究 优先出版	左青云;陈鸣;赵广松;邢长友;张国敏;蒋培成	软件学报	2013/05	267	7592		
2	基于多类特征的Android应用恶意行为检测系统	杨欢;张玉清;胡予濮;刘奇旭	计算机学报	2014/01	39	2192		
3	社会网络数据发布隐私保护技术综述 优先出版	刘向宇;王斌;杨晓春	软件学报	2014/03	33	1819		

图13-32 CNKI中国学术期刊网分类检索实例

从图 13-32 可以得出在“网络安全”方面的学术论文总量为 75 819 篇(截至 2016 年 7 月),学术价值较高且排名靠前的学术论文的作者、期刊名、被引量、下载量等有价值的信息,对于进一步下载和阅读全文内容有着重要参考价值。

(2) CNKI 中国学术期刊网一般检索。它就是直接用主题、篇名、关键词、作者、作者单位、刊名、ISSN、CN、期、基金、摘要、全文、参考文献、中图分类号、DOI、栏目信息 16 种期刊论文信息字段进行检索。默认输入两个检索词,可以根据需要应用“+”和“-”增删检索词输入的最大量,利用“+”可以增加最多的检索项为 14 个。同时可以控制信息的来源类别,默认为全部期刊。一般检索视图见图 13-33。

图 13-33　CNKI 中国学术期刊网一般检索视图

(3) CNKI 中国学术期刊网高级检索。可以对最多 14 个检索词进行布尔逻辑组配,同时可以对时间段、更新时间(最近半年、最近一月等)、期刊来源、支持基金、作者与作者单位等检索项进行限定。高级检索视图见图 13-34。

图 13-34　CNKI 中国学术期刊网高级检索视图

(4) CNKI 中国学术期刊网专业检索。专业检索用于图书情报专业人员查新、信息分析等工作,使用逻辑运算符和关键词构造检索式进行检索。跨库专业检索支持对以下检索项的检索:SU='主题',TI='题名',KY='关键词',AB='摘要',FT='全文',AU='作者',FI='第一责任人',AF='机构',JN='中文刊名'&'英文刊名',RF='引文',YE='年',FU='基金',CLC='中图分类号',SN='ISSN',CN='统一刊号',IB='ISBN' ,CF='被引频次'。

“AND”、“OR”、“NOT”三种逻辑运算符的优先级相同;如要改变组合的顺序,请使用英文半角圆括号“()”将条件括起;逻辑关系符号(与(AND)、或(OR)、非(NOT)前后要空一个字节;使用“同句”、“同段”、“词频”时,需用一组英文单引号将多个检索词及其运算符括起,如'流体#力学'。假设检索钱伟长在清华大学或上海大学期间发表的文章。检索式:AU=钱伟长 and (AF=清华大学 or AF=上海大学)。假设要求检索钱伟长在清华大学期间发表的题名或摘要中都包含“物理”的文章,检索式:AU=钱伟长 and AF=清华大学 and (TI=物理 or AB=物理),实例如图 13-35 所示。

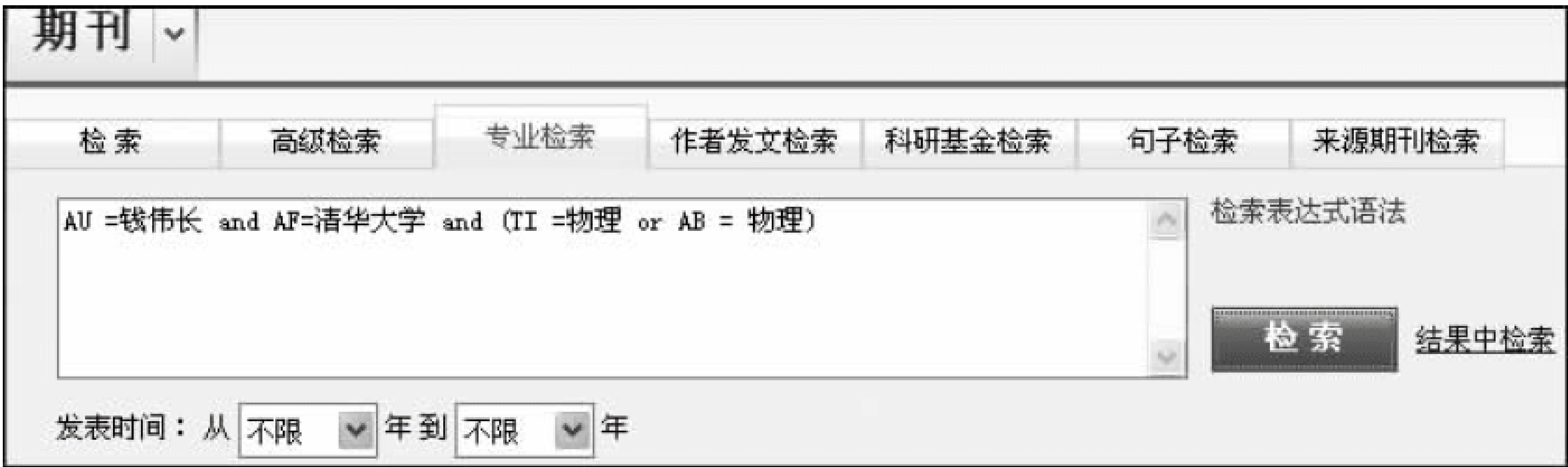


图 13-35 CNKI 中国学术期刊网专业检索实例

(5) CNKI 中国学术期刊网作者发文检索。为了追踪某一专家学者的学术成果(有些专家一生可能在多个单位工作过),以便于发现其研究动向,横向或纵向比较同领域研究者的学术动态,“作者发文检索”则提供了有益的辅助功能。见图 13-36。

(6) CNKI 中国学术期刊网科研基金检索。一般学术研究(包括基础性研究或应用性研究)都受到一定机构或不同级别的专门科研基金资助,以保障研究项目与项目任务的顺利完成,因此可以用“科研基金”途径检索学术论文,检索时直接输入基金名即可。在不清楚具体基金名称时,可利用基金分类目录查询。见图 13-37。

(7) CNKI 中国学术期刊网句子检索与来源期刊检索。在检索学术论文的全文时,如果没有明确的主题词和关键词作为检索项,可以用句子(或一句话)作为整体检索项。句子检索最多可容纳四个句子,在全文中是否在“同一句”或“同一段”。见图 13-38。

来源期刊检索是依据学术论文发表和登载的具体期刊类型来筛选论文信息,默认为

图 13-36　CNKI 中国学术期刊网作者发文检索视图

图 13-37　CNKI 中国学术期刊网基金分类目录检索视图

图 13-38　CNKI 中国学术期刊网句子检索视图

全部期刊。例如，限定来源期刊类别为“SCI 来源期刊”和“CSSCI”，也就大致确定了结果论文的等级与参考价值。见图 13-39。

13.3.2　维普中文科技期刊数据库检索

维普《中文科技期刊数据库》（简称维普期刊数据库）是由国家科技部西南中心研制开发的我国第一个海量期刊数据库，它主要收集我国公开或非公开发行的各种期刊，该库已



图 13-39　CNKI 中国学术期刊网来源期刊检索视图

经成为我国数字图书馆建设的核心资源之一，是高校图书馆文献保障系统的重要组成部分，也是高校师生、科研工作者进行科技查证和科技查新的常用数据库。该数据库涵盖期刊总数为 12 000 余种，其中核心期刊 1810 种，目前提供服务的全文文献总量达 5000 多万篇，数据更新周期为每周一次，每年的数据增量达 300 万篇。采用国际通用的高清晰 PDF 全文数据格式处理数据，为读者提供八大类学术期刊论文服务，即社会科学、自然科学、工程技术、农业科学、医药卫生、经济管理、教育科学和图书情报。

(1) 维普期刊数据库的一般检索。直接输入检索词(主题词、关键词、作者、刊名等)即可。一般检索界面见图 13-40。



图 13-40　维普期刊数据库的一般检索界面

(2) 维普期刊数据库高级检索。可以使用最多五个检索项进行逻辑组配检索，同时可以限定期刊论文的时间段、更新时间和来源期刊范围(例如核心期刊、SCI 来源期刊等)，以提高学术论文检索的返回结果精度。高级检索界面见图 13-41。

(3) 维普期刊数据库专业检索式检索。专业检索式检索与 CNKI 期刊数据库的原理相似，检索是 AND 代表“并且”，OR 代表“或者”，NOT 代表“不包含”(注意必须大写，运算符两边需空一格)。例如，需要 C++ 或 Basic 方面且是计算机应用与软件属性的信息，但是 Visual 方面的除外，则检索表达式为：J=计算机应用与软件 AND (U=C++ OR U=Basic) NOT M=Visual。见图 13-42。

高级检索 检索式检索

M=题名或关键词 模糊
与 A=作者 模糊
或 I=基金资助 模糊
非 S=机构 模糊
与 J=刊名 模糊

时间限定

时间：1989 - 2016　更新时间：三个月内

期刊范围

全部期刊 核心期刊 EI来源期刊 SCI来源期刊 CAS来源期刊 CSCD来源期刊 CSSCI来源期刊

图 13-41　维普期刊数据库高级检索界面

高级检索 检索式检索

检索规则说明：AND代表"并且"；OR代表"或者"；NOT代表"不包含"；(注意必须大写,运算符两边需空一格)

检 索 范 例 ：范例一：(K=图书馆学 OR K=情报学) AND A=范并思　范例二：J=计算机应用与软件 AND (U=C++ OR U=Basic) NOT M=Visual

J=计算机应用与软件 AND (U=C++ OR U=Basic) NOT M=Visual

时间限定

时间：1989 - 2016　更新时间：一个月内

期刊范围

全部期刊 核心期刊 EI来源期刊 SCI来源期刊 CAS来源期刊 CSCD来源期刊 CSSCI来源期刊

图 13-42　维普期刊数据库专业检索式检索界面

13.4　典型外文电子图书检索系统

13.4.1　CADAL 外文图书检索

大学数字图书馆国际合作计划(China Academic Digital Associative Library, CADAL)前身为高等学校中英文图书数字化国际合作计划(China-America Digital Academic Library,CADAL)。项目由国家投资建设,作为教育部“211”重点工程,由浙江大学联合国内外的高等院校、科研机构共同承担。项目负责人为浙江大学潘云鹤院士。CADAL 一期建设完成 100 万册(件)数字资源,CADAL 二期建设完成 150 万册(件)数字资源,包括外文图书 55 万册,系统服务网址为:http://www.cadal.cn/。

在检索时可以用“搜全部”、“仅搜书名”、“仅搜作者”和“搜索词完全匹配”四种形式进行检索限定。例如用“from china”作为检索词进行模糊查询,可以检索到相关书籍 683 种,实例如图 13-43 所示。



图 13-43　CADAL 外文图书检索实例

读者可以在检索结果排序中选择需要阅读的“图书封面图标”,直接在线阅读电子图书全文内容或者借阅纸质印刷版图书。

13.4.2　世界电子图书馆检索

世界电子图书馆(World eBook Library,WeL)是世界公共图书馆联盟(World Public Library Association,WPLA)的电子图书项目,WPLA 成立于 1966 年,网址为:http://www.ebooklibrary.org/,是非营利性的世界组织。世界电子图书馆是世界最大的电子书提供商,不属于任何机构或者部门,资源收集来源于世界 20 万家出版机构的电子文献。

世界电子图书馆的资源内容覆盖了 31 个学科大类,共计 152 个学科种类(如文学、历史、政治、社会学、教育、经济、法律、戏剧等学科),以人文社会科学为主,还包括自然科学、

农学、医学、工程技术等领域的经典文学作品、书籍、期刊、百科全书、字典、手册等参考资源，共有全球 260 多种语言的超过 340 万册 PDF 格式电子图书与 23 000 多种有声读物。WeL 外文电子图书普通检索视图见图 13-44。

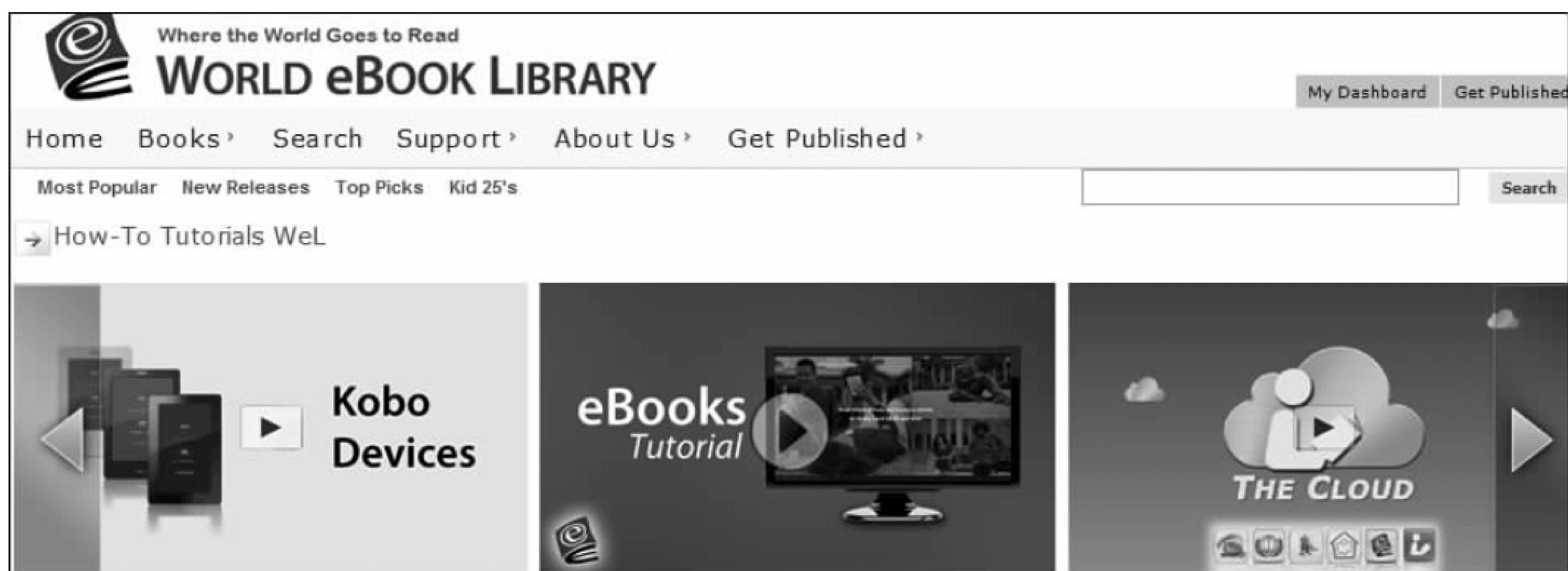


图 13-44　WeL 外文电子图书普通检索视图

WeL 特色专集值一：Graphic Novel Collection 图像小说专集。《大英百科全书》对图像小说的定义是：图像小说，在美国和英国的用法中，指一种联合了文字与图片——漫画图片的文本。对比“漫画(Comic)”，“图像小说通常是指针对成年读者的长篇漫画叙事，以精装或者平装书籍出版并用过书店销售，它探讨严肃的文学性主题，并且绘制精美”。WeL 的 Graphic Novel Collection 中主要包括三种类型的资源，而不仅限于上述严格意义上的图像小说：针对年轻读者的插图小说、科幻冒险图像小说(Sci-Fi and Adventure)及漫画(Comic)，资源数量超过 11 000 册。

WeL 特色专集之二：经典原著。2005 年，世界上最大的提供文献信息服务的机构之一 OCLC(Online Computer Library Center, Inc.，即联机计算机图书馆中心)通过对其 56 000 家联盟成员馆的馆藏资源进行调查分析，得出了一份“TOP 1000”图书名单，这份名单上的图书资源被认为是值得世界上所有图书馆收录的永恒经典(“Timeless Classics”)。WeL 中包含了列表中 70%以上的图书资源，同时还提供超过 10 000 种类似的图书供阅读。

WeL 特色专集之三：创新。科学研究的目的在于通过科学研究“发现”或者“创造”，以达到“改变”世界的目的，即“创新”，创新不仅是科学研究的目的，也是其灵魂。科学研究成果的创新性内容是其重要的评价标准。WeL 数据库拥有超过 400 余种“Innovation”相关电子图书与 11 000 余种“Innovation”相关文献。

WeL 特色专集介绍之四：参考工具书。WeL 中收录了上万本的参考工具书(字典、词典、传记、百科全书、手册等)，例如《大英百科全书》(*Encyclopedia Britannica*)、《汉英双语词典》(*Chinese-English Dictionary*)、《布莱恩画家和雕刻家词典》(*Bryan's Dictionary of Painters and Engravers*)等。

WeL 支持快速检索与高级检索两种检索方式，也可选择系统推荐的检索表达式快速检索到所需资源。检索结果按照作者、学科、出版社、语言、文件格式以及专题进行聚类，可快速定位所需文献。

WeL 外文电子图书高级检索包括的字段有所有字段、题名、作者、学科和出版社，同时可以限制出版时间、图书语言、文件格式、学科分库、主题分库等图书范围，支持关键词检索、精确检索和逻辑检索。高级检索视图见图 13-45。

Find books that have...
Everything: [All These Words]
Title: [All These Words]
Author: [All These Words]
Subject: [All These Words]
Publisher: [All These Words]
All These Words
Exact Phrase
Any of These Words
Except These Words
Year of Publication Between: and
Language: Select Languages...
File Type: Select File Types...
Academic Collection: Select Academic Collection...
eBook Library Collection: Select eBook Library Collections...

图 13-45　WeL 外文电子图书高级检索视图

在校园网内的大学生用户，可以无限制地阅读、下载甚至可以非商业目的打印整本电子书，且所下载的电子图书文件可永久保存；读者在 WeL 中注册后，可使用系统的“我的阅读历史”、“我的书单”、“上传电子图书”等个性化服务功能；通过社区功能，可以阅读和分享书评、添加评论等。所有电子资源都采用 PDF/Mp3/Mp4 格式，安装 PDF 阅读器(如 Adobe Reader)与 Mp3/Mp4 播放器即可打开资源内容。

13.4.3　ebrary(电子图书馆)检索

ebrary 公司于 1999 年 2 月正式成立，由 McGraw-Hill Companies、Pearson plc 和 Random House Ventures 三家出版公司共同投资组建。ebrary 电子图书数据库整合了来自 500 多家学术、商业和专业出版商的权威图书和文献，覆盖商业经济、社科人文、历史、

法律、计算机、工程技术、医学等多个领域。截至 2015 年 3 月，ebrary 的综合学术类收藏(Academic Collection)中已包含了 12.4 万多册图书。

登录个人账号之后，才可按章节下载，或按页码下载，还可以下载全文，所下载的全文需要使用 Adobe Digital Editions 工具阅读。未注册用户可单击页面右上角的“Sign in”，再单击“Create an account”，自行注册设置用户名和密码。

ebrary 是一个高度交互式的电子图书集合，阅读 ebrary 资源需要下载 ebrary Reader 专门阅读器。图 13-46 是用关键词 database 进行检索，返回结果集为 50 426 个电子书文档。

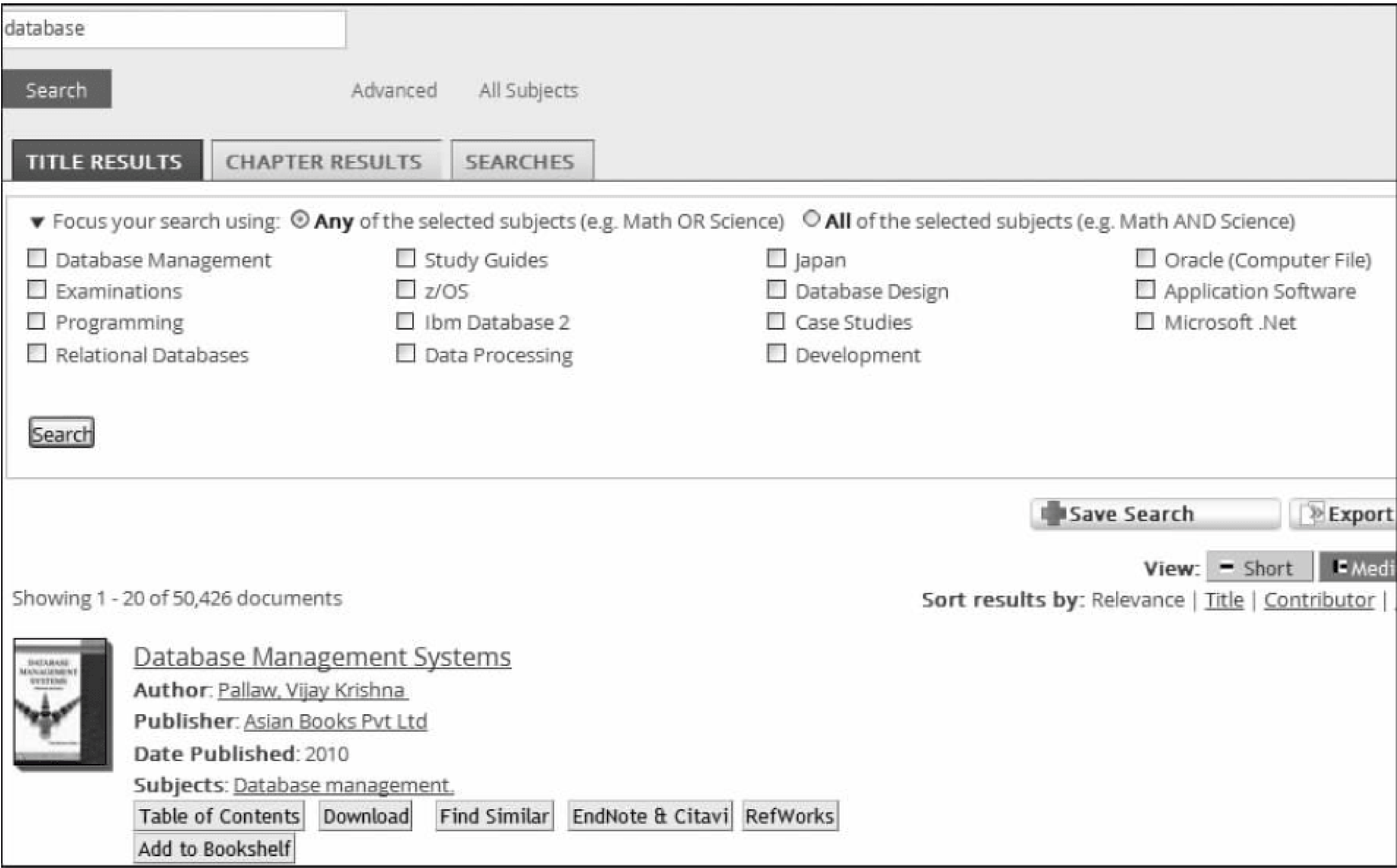


图 13-46　ebrary 普通检索实例

对于需要的图书可以开展在线阅读、下载、添加到书架、近似查找、阅读批注、参考查阅等相关学习行为。

ebrary 高级检索通过“＋”和“－”控件来调节检索词的数量，高级检索最多可以容纳九个检索项，且各个检索项之间在关系表达上为“逻辑与”。图 13-47 是用 database、management 和 design 为检索词进行的高级检索实例(要求 database、management 出现在书名 title 中，design 出现在主题 subject 中)。

图 13-47　ebrary 高级检索实例

13.4.4　OCLC FirstSearch 检索

OCLC(Online Computer Library Center, Inc.),即联机计算机图书馆中心,总部设在美国的俄亥俄州,是世界上最大的提供文献信息服务的机构之一,它是一个非营利的组织,以推动更多的人检索世界上的信息、实现资源共享并减少使用信息的费用为主要目的。OCLC 的 FirstSearch 是一个面向最终用户设计的交互式联机信息检索系统。其通用检索视图见图 13-48。

由于 OCLC 存储有海量数据,需要用户在输入检索词时,同时要选择具体数据库,然后再检索,这样会大大减少用户对结果的评价工作量。进行高级检索时,可以跨库检索,最多选择两个数据库;也可以用多个检索词进行逻辑组配,构造需要的检索表达式,进行专家检索。

例如,(au: Shak * not au: Shakespeare) not mt: juv and yr: 1999 and dt=“bks”的含义是:著者名字以“Shak”开头,但不是“Shakespeare”,不归类于期刊,年份是 1999 年,限制类型是图书。



图 13-48　OCLC FirstSearch 通用检索视图

OCLC FirstSearch 可检索的主要图书数据库有以下几种。

(1) EBooks 电子书书目数据库。Ebooks 收录了 OCLC 成员图书馆编目的所有电子书的书目信息，接近 1300 万种，涉及所有主题，涵盖所有学科，收录日期从公元前 1000 年至今。数据更新频率为每天。

(2) GPO 美国政府出版物书目数据库。GPO 美国政府出版物数据库(U. S. Government Printing Office)由美国政府出版署创建，覆盖从 1976 年以来各种各样的美国政府文件，包括美国国会的报告、听证会、辩论、记录、司法资料以及由行政部门(国防部、国务院、总统办公室等)颁布的文件，每条记录包含有一个书目引文，共有 60 多万条记录。数据更新频率为每月。

(3) WorldCat 联机联合目录数据库。WorldCat 联机联合目录数据库是世界上最大的书目记录数据库，包含 OCLC 近两万家成员馆编目的书目记录和馆藏信息。从 1971 年建库到目前为止，共收录有 480 多种语言总计达 20 亿多条的馆藏记录、3 亿多条独一无二的书目记录，每个记录中还带有馆藏信息，基本上反映了从公元前 4800 多年至今世界范围内的图书馆所拥有的图书和其他资料，代表了四千年来人类知识的结晶。文献类型多种多样，包括图书、手稿、地图、网址与网络资源、乐谱、视频资料、报纸、期刊与杂志、文章以及档案资料等。该数据库平均每十秒更新一次。

13.4.5 其他典型外文电子图书检索系统简述

(1) Early English Books Online(早期英文图书在线,简称 EEBO)。它是由密歇根大学、牛津大学和 ProQuest 公司合作开发并于 1999 年推出的在线全文数据库。EEBO 收录了 1473—1700 年之间所有现存的英语世界出版物资料,其中包括许多知名作家的著作,例如莎士比亚(Shakespeare)、马洛礼(Malory)、斯宾塞(Spencer)、培根(Bacon)、莫尔(Moore)、伊拉斯漠(Erasmus)、鲍尔(Bauer)、牛顿(Newton)、伽利略(Galileo)等。除了收录那个时期的大量文学资料外,EEBO 还收录许多历史资料,例如皇家条例及布告、军事、宗教和其他公共文件,年鉴、练习曲、年历、大幅印刷品、经书、单行本、公告及其他的原始资料。EEBO 覆盖历史、英语文学、宗教、音乐、美术、物理学、妇女问题研究等诸多领域。

(2) iG Publishing 电子图书。iG Publishing 电子图书包括以下九个电子图书数据库中的图书、工具书,都由行业中权威学会或出版社出版。读者可直接在全部数据库中一并检索,也可分别进入具体的九个数据库中浏览及检索图书。

① 美国材料信息学会(ASM International)电子图书数据库:美国材料信息学会自 1913 年成立以来,一直致力于材料科学和工程专业的研究发展。

② 英国标准学会(The British Standards Institution)电子图书(手册)数据库:英国标准学会是世界上第一个国家标准化机构,成立于 1901 年,总部设在伦敦。

③ 国际工程联合会(International Engineering Consortium,IEC)电子图书数据库:美国国际工程联合会成立于 1944 年,最初由美国各大学和工程组织联合发起,专注于电子工业的再教育。

④ 美国工业出版社(Industrial Press)电子图书数据库:从 1883 年成立以来,美国工业出版社一直恪守其出版传统,以最好的技术为教育事业提供优秀的参考书。

⑤ 美国摩根出版社(Morgan & Claypool Publishers)电子图书数据库:出版社成立于 2002 年,其出版的综述文集(Synthesis)为工程、计算机科学、生命科学领域及相关领域(如材料、能源、环境等)的研发和教育工作者提供了一种创新型的信息服务。

⑥ 英国多科学出版有限公司(Multi Science)电子图书数据库:公司成立于 1961 年,其出版物包括三个学科领域:能源、声学和工程科学。

⑦ 英国皇家建筑学会(Royal Institute of British Architects,RIBA)电子图书数据库:RIBA 是英国建筑机构和建筑行业的专家,他们为其成员提供世界范围内的各种培训形式、技术服务,以及出版物和活动,并为在英国和海外的建筑师教育设定了标准。

⑧ 美国科学技术出版社(SciTech Publishing)电子图书数据库：美国科学技术出版社已经在雷达和国防电子学领域成为了全球出版领导者。

⑨ 美国工业和应用数学学会(Society for Industrial and Applied Mathematics, SIAM)电子图书数据库：SIAM 是一个以促进应用和计算数学的研究、发展、应用为目的的协会。

(3) Knovel。Knovel 是一个统一的信息平台，具有强大的检索和分析功能，目前收录了来自 120 多个出版机构的实践经验、验证的方程和材料及物质数据，可以帮助用户快速找到解决技术问题的答案。Knovel 将工程学和应用科学的数据信息与分析、检索工具整合在一起，提供“交互式”的数据分析功能，从而让数据表格及图表“活”了起来。

(4) MyiLibrary。MyiLibrary 电子图书平台在世界范围内合作的出版商超过 400 家，其中包括世界著名的学术出版商和出版社，如 Taylor&Francis，Wil Balckwell，Oxford University Press，Cambridge University Press 等。该平台目前包含有电子书 12 000 多种，涉及教育、艺术、法律、文学、医学、哲学、心理学、政治学、工程技术、自然科学、图书馆学等领域。

该平台上还包括培生教育出版集团(Pearson Education Group)出版的 982 种电子教材全文，内容涉及数学、物理、化学、工程、计算机科学、信息技术、生物学、心理学、社会学、法律、商业管理、经济、市场营销、金融、教育、就业指导、英语、艺术等学科。培生教育出版集团是目前全球最大的教育出版集团，这些电子教材是该集团为教育部外国教材中心特别提供的。平台上所有的电子书可进行全文检索；还可按关键词、作者、ISBN、出版年、学科、语种等对检索结果进行限定。

(5) Safari。Safari 由世界两大著名 IT 出版商 O'Reilly & Associates，Inc. 和 The Pearson Technology Group 共同组建，主要提供 IT 类的电子图书，其中，95%以上是 2000 年以后出版的，22%的书目列入了 Amazon 书店前 10 000 种需要的图书清单中。

Safari 覆盖的主题包括 Programming、Operating Systems、Networking 等。在 Safari 中可以按主题或出版商分类浏览图书，可进行高级检索，并可直接定位浏览书中的编程信息。阅读全文时可由检索结果中的“Table of Contents”直接跳到书中章或节，也可单击图书封面，再选择页面右侧的“Start Reading”从头开始阅读。

(6) Wiley Online Library。John Wiley & Sons Inc. 是有 200 多年历史的国际知名专业出版机构，在化学、生命科学、医学以及工程技术等领域学术文献的出版方面颇具权威性，2007 年 2 月与 Blackwell 出版社合并，两个出版社的出版物整合到同一平台上提供服务。Wiley Online Library 是一个综合性的网络出版及服务平台，在该平台上提供全文

电子期刊、在线图书 、在线参考工具书以及实验室指南。

(7) World Bank E-library。可以在线阅读世界银行所有有关社会和经济类的全文图书、报告和多种文件。它带给读者的是一个全文检索和多重查询的数据库。迄今为止,该在线图书馆已提供了世界银行从1987年以来出版的4500多种图书、所有世界银行政策研究工作报告和各种文件的全文内容,同时介绍即将出版的图书信息等。每年新增150～175本图书,新增250～300个工作报告。

13.5 典型外文学术期刊检索系统

13.5.1 Web of Science 数据库检索

Web of Science 数据库收录了12 400多种世界权威的、高影响力的学术期刊,内容涵盖自然科学、工程技术、生物医学、社会科学、艺术与人文等领域的海量学术研究论文,最早回溯至1900年。Web of Science 收录了论文中所引用的参考文献,并按照被引作者、出处和出版年代编制成独特的引文索引。

1. Web of Science 数据库的主要构成

Web of Science 是获取全球学术信息的重要数据库,由以下几个重要部分组成。

(1) Science Citation Index－Expanded (SCIE,科学引文索引)。

(2) Social Sciences Citation Index (SSCI,社会科学引文索引)。

(3) Arts & Humanities Citation Index (A&HCI,艺术人文引文索引)。

(4) Conference Proceedings Citation Index (CPCI,会议论文引文索引)。

(5) Current Chemical Reactions 收录了1840年以来的化学反应的事实性数据。

(6) Index Chemicus 收录了1993年以来的化学物质的事实性数据。

2. Web of Science 数据库检索与利用的主要作用

Web of Science 作为全球权威的引文数据库,广泛收录了世界一流的学术研究成果。其强大的分析功能,更能够在快速锁定高影响力论文、发现国内外同行权威所关注的研究方向、揭示课题的发展趋势、选择合适的期刊进行投稿等方面帮助研究人员更好地把握相关课题,寻求研究的突破与创新点。

(1) 随时掌握课题的最新进展。

(2) 了解相关领域中最具影响力的研究人员。

(3) 对著作中重要理论的发展和应用进行跟踪。

(4) 选择合适的学术期刊发表论文。

(5) 寻找合作研究者或深造机会。
(6) 准确查找论文的被引用情况。
(7) 按照所投稿期刊的格式快速生成参考文献。
(8) 在网络平台上建立个人图书馆。

3. Web of Science 基本检索

所有成功的检索均添加至检索历史表。在创建检索式时,需要遵循所有适用的检索规则。可以在"检索"页面中最多选择三个字段作为默认检索字段。在检索式中最多可输入 6000 个检索词。添加新的字段还会将第二个字段设置为 AND 运算符,可以将 AND 运算符改为 OR 或 NOT。用于检索的基本字段有主题、标题、出版物名称、作者、编者、出版年等。基本检索视图见图 13-49。

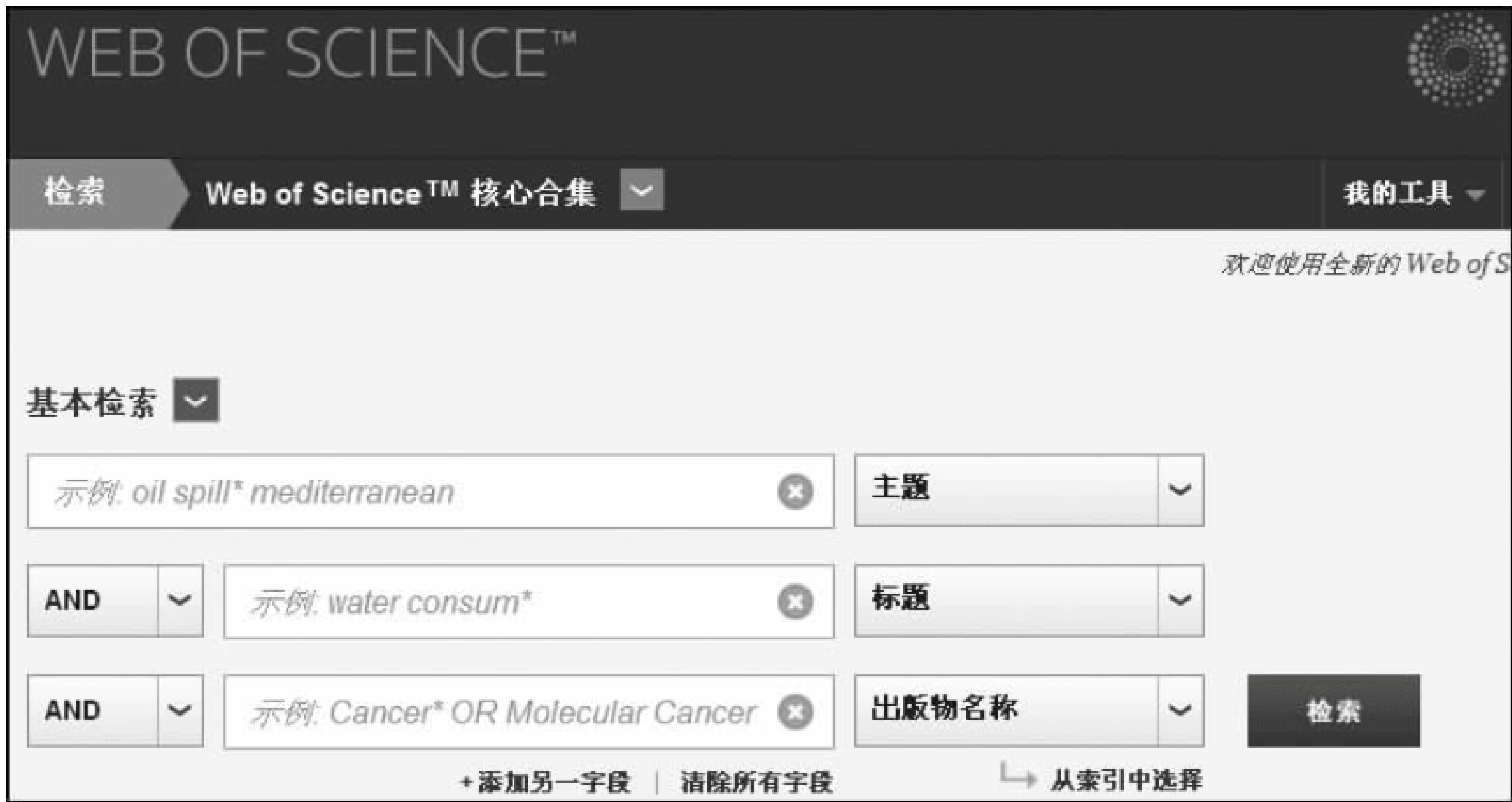


图 13-49　Web of Science 基本检索视图

在基本检索时,默认检索字段数为 1,用户随时可以使用"添加另一字段"添加更多的检索字段,或者可以从"检索"页面删除检索字段。一个检索字段:默认字段始终为"主题",随时可以选择不同的检索字段。三个检索字段:默认字段始终是"主题"、"作者"和"出版物名称"。添加另一字段:默认字段始终为"主题",随时可以选择不同的检索字段。

基本检索方法如下。

(1) 在大多数字段输入两个或两个以上相邻的检索词时,产品会使用隐含的 AND。例如,在"主题"或"标题"检索时输入 rainbow trout fish farm 与输入 rainbow AND trout AND fish AND farm 是等效的,这两个检索式会返回相同数量的检索结果。

(2) 如果要更改检索设置(包括不同数据库选择),请转至检索页面的时间跨度和更

多设置部分。

(3) 在一个或多个检索字段中输入检索词。在执行检索时,也可以使用如下选项。

① 添加另一字段链接用于向“基本检索”页面添加更多的检索字段。

② 重置表单链接用于清除已输入的任何检索式。此操作将检索页面重置为原始检索字段,适用于“作者”检索和“被引参考文献”检索。

③ 从索引选择链接用于在执行“出版物名称”或“作者”检索时选择一个项目。

④ 自动建议的出版物名称选项用于打开或关闭出版物名称的自动建议。当开启此功能时,产品根据用户在检索字段中输入的字符提供出版物名称的列表。例如,如果您输入 CANC,则产品显示以这四个字符开头的出版物列表,如 Cancer Biology Therapy 和 Cancer Investigation。

⑤ 显示的默认检索字段数选项允许仅选择“主题”字段,或者可以选择“主题”、“作者”和“出版物名称”字段。保存设置选项用于保存您的设置以供将来的检索会话使用。

4. Web of Science 高级检索

基本方法是在每个检索式编号前输入数字符号(#),检索式组配中包括布尔运算符(AND、OR、NOT),使用括号可以改写运算符优先级。如表 13-1 所示。

表 13-1 Web of Science 高级检索举例

检索式	检索结果	完整检索式
#3	727	#2 AND #1
#2	1 125 241	AD=(Japan OR Russia)
#1	31 082	TI=(cell death OR apoptosis)

表 13-1 中,检索式#3 找到的记录在“标题”中出现 cell death 或 apoptosis,并且“地址”字段中出现 Japan 或 Russia。

Web of Science 高级检索类似于其他检索系统的“专业检索”,需要对复杂检索需求进行逻辑检索式构造。主要字段标识:AD=地址,AI=作者标识符,AU=作者,CF=会议,CI=城市,CU=国家/地区,DO=DOI,ED=编者,FG=授权号,FO=基金资助机构,FT=基金资助正文,GP=团体作者,IS=ISSN/ISBN,OO=组织,PY=出版年,SO=出版物名称,SU=研究方向,TI=标题,TS=主题,WC=Web of Science 类别。高级检索实例见图 13-50。

在某一研究方向领域检索。在高级检索中使用 SU 字段标识以查找研究方向检索



图 13-50　Web of Science 高级检索实例图

词,将检索范围缩小至特定研究领域。例如,SU=(Biochemistry & Molecular Biology AND Biophysics)可查找在全记录的“研究方向”字段中同时出现检索式里的这两个研究方向的记录。

在某一类别方面检索。使用 Web of Science 类别以及高级检索的 WC 字段标识,将检索范围缩小至特定研究领域。例如,WC=(Anthropology AND Archaeology)可查找在全记录的“Web of Science 类别”字段中同时出现检索式里的这两个类别的记录。

13.5.2　IEL 数据库检索

IEL 的全称为 IEEE/IET Electronic Library,它是 IEEE 旗下最完整、最有价值的在线数字资源,通过智能的检索平台(http://ieeexplore.ieee.org/Xplore)为用户提供创新的文献信息。其权威的内容覆盖了电气电子、航空航天、计算机、通信工程、生物医学工程、机器人自动化、半导体、纳米技术、电力等各种技术领域。IEL 数据库提供 IEEE(电气电子工程师学会)和 IET(国际工程和技术学会)出版的以下几类刊物的全部资源。

(1) 170 余种 IEEE、20 余种 IET 期刊与杂志、1 种 BLTJ 期刊,总数达 400 多种(包括过刊及更名刊)。

(2) 每年 1400 多种 IEEE 会议录和 20 多种 IET 会议录,总量超过 17000 卷。

(3) 60 多种 VDE 会议录,超过 4500 篇。

(4) 2600 多种 IEEE 标准(包括现行标准和存档标准,标准草案需额外订购)。

(5) 390 多万篇全文文档,提供 1988 年以后的全文文献,部分历史文献回溯到

1872 年。

IEL 每月增加 25 000 篇最新文献,且每年 IEEE 还有新的出版物加入到 IEL 中。据 ISI 每年的期刊引用报告,IEEE 连续高居众多技术领域的前列。IEEE 出版物是电气和电子工程领域最重要的文献资料,约占全世界该领域核心文献的 30%。

1. IEL 数据库基本检索

直接输入检索的主题词或关键词(例如 page ranking),如果是学校的校园网用户,无须注册可直接检索,同时在检索界面的顶部正中央会出现高校名称,如图 13-51 所示来自"桂林电子科技大学"(GuiLin University Of Electronic Technology)的合法用户。

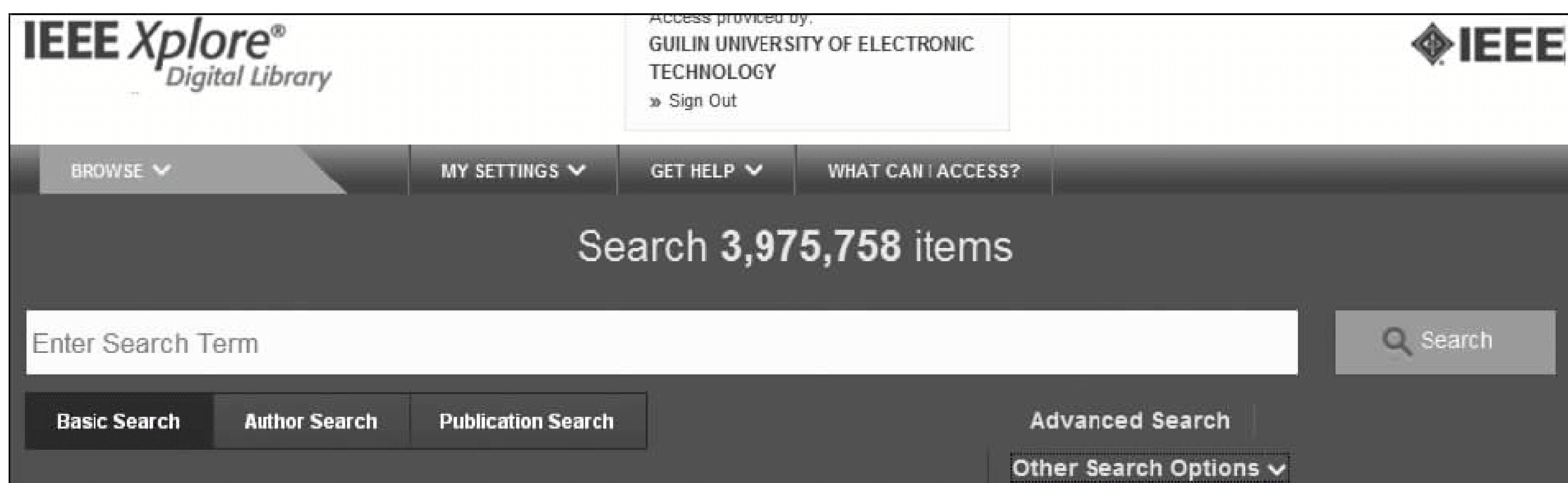


图 13-51 IEL 数据库基本检索实例

在基本检索(basic search)界面中,显示了当前的信息资源总量为 3 975 758 项(截止到 2016 年 7 月),同时表明检索途径丰富,包括作者检索、出版项检索、高级检索以及其他检索选择等。

在资源类型与范围方面,展开 BROWSE 可以选择其中任意一种资源类型:图书与电子书、会议出版物、课程、期刊杂志、标准和热点导航。

2. IEL 数据库著者检索与出版项检索

著者检索可以用著者的家族名(family name)、姓(last name)、名(surname)。著者检索可以聚类考查某一学者或专家的总体研究情况和最新研究趋势,也便于与其合作与交流。出版项检索需要用出版物的卷(volume)、期(issue)或开始页(start page)进行检索。

3. IEL 数据库高级检索

IEL 数据库高级检索包括三种类型:关键词或短语检索、命令检索和索引检索。关键词或短语检索类似于中文数据库的关键词与主题检索,命令检索类似于专业检索。

(1) 关键词或短语检索。第一,输入关键词或短语(默认为两个),可以实际检索需要通过 Add New Line 增加检索输入项,也可以通过删除按钮或 Reset All 来调整检索项数

量。第二,选择资源范围,包括元数据(在字段中查询)、全文或元数据(可以模糊查询全文内容)。第三,选择多个关键词与短语之间的逻辑运算关系(AND、OR、NOT)。第四,指定每个检索词的字段项(Authors、ISSN 等)。IEL 数据库的高级关键词与短语检索视图见图 13-52。



图 13-52　IEL 数据库的高级关键词与短语检索视图

(2) 命令检索。Command Search 主要采用比较规范的检索命令用逻辑运算符(AND、OR、NOT、NEAR、ONNEAR)将检索项组配起来,构成一致的检索表达式。IEL 数据库的命令检索实例见图 13-53。

IEL 数据库的命令检索形式,例如,Abstract": ofdm AND " PublicationTitle": communications;"Author":"Suzuki, T";(java or XML) AND "software engineering"; security NEAR/5 "cloud computing";"Fast" ONEAR/5 "Statistic" AND "Document Title": "Fast";(("Abstract": java) OR "Publication Title": "computer technology") AND "Document Title": rfid。IEL 数据库命令检索的更多形式可以进一步参考链接 http://ieeexplore.ieee.org/Xplorehelp/Help_searchexamples.html。

(3) 索引检索。可以直接在数字对象唯一标识符(digital object unique identifier)输

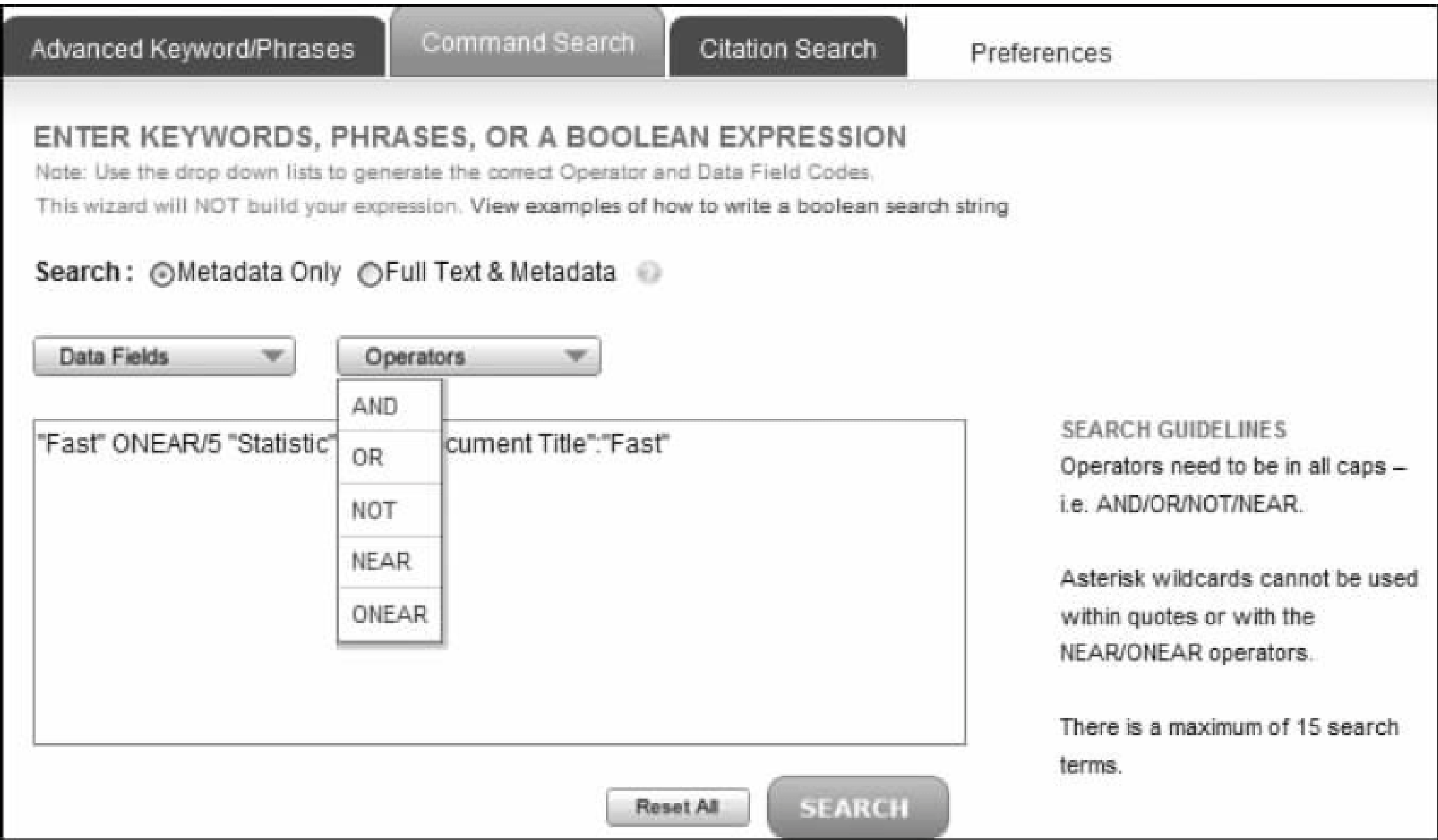


图 13-53　IEL 数据库的命令检索实例

入唯一的检索号，也可以在 title 中输入关键词或短语查询。通过索引查询一定主题，可以洞察其研究脉络、研究者之间的相互影响关系及其研究趋势，这对于自身的研究探索与研究创新有很好的参考价值。

（4）检索举例。在高级检索中用 network、security、algorithm 为检索词，它们之间用逻辑与“AND”关系，且 network、security 出现在 Document Title 位置，进行检索。见图 13-54。

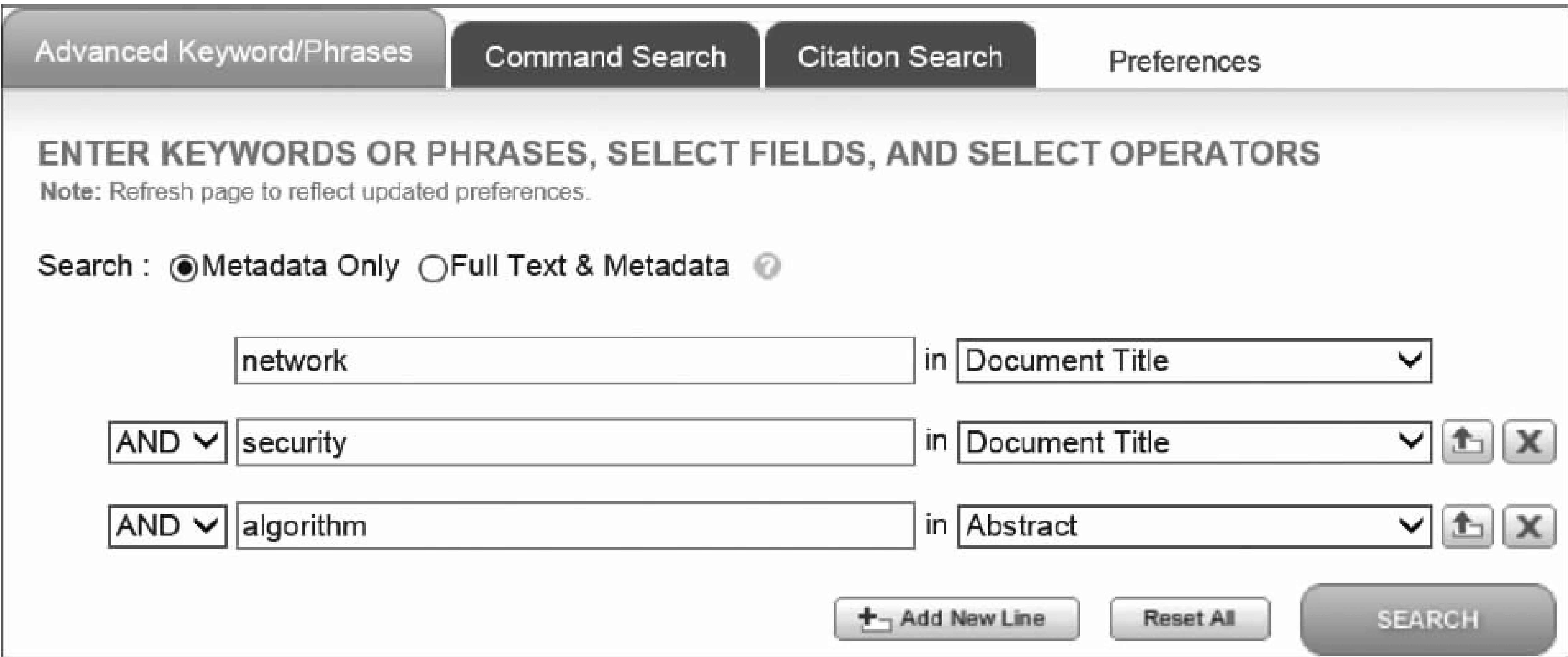


图 13-54　IEL 数据库高级检索实例一

然后获得如下检索结果共 542 项所需信息即“网络安全算法”方面的学术论文，使用

sort by 对检索结果数据进行排序(包括相关性排序、最新更新排序、最高被引排序等)。为了参考相关信息,一般选择“Most Cited”排序,查看一些学术价值高的论文。见图 13-55。

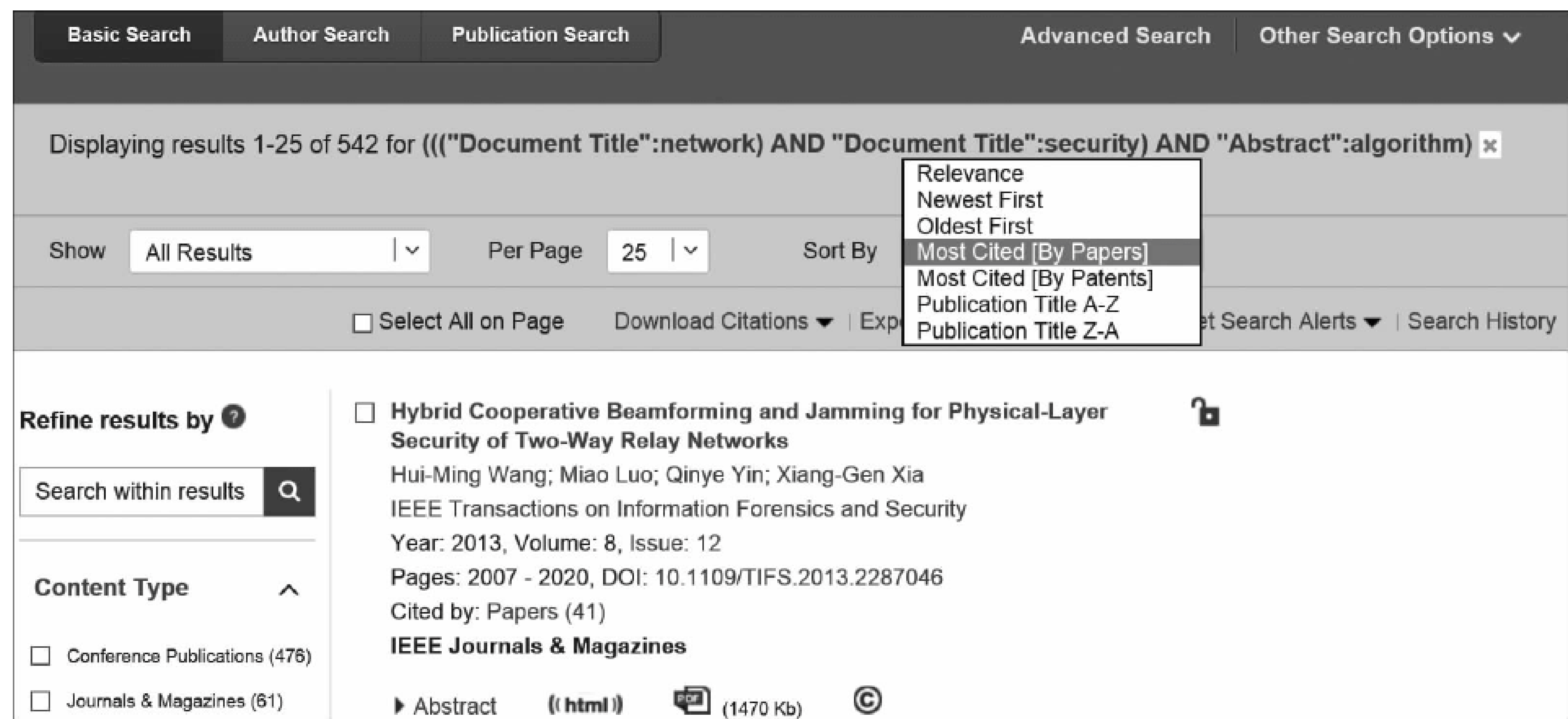
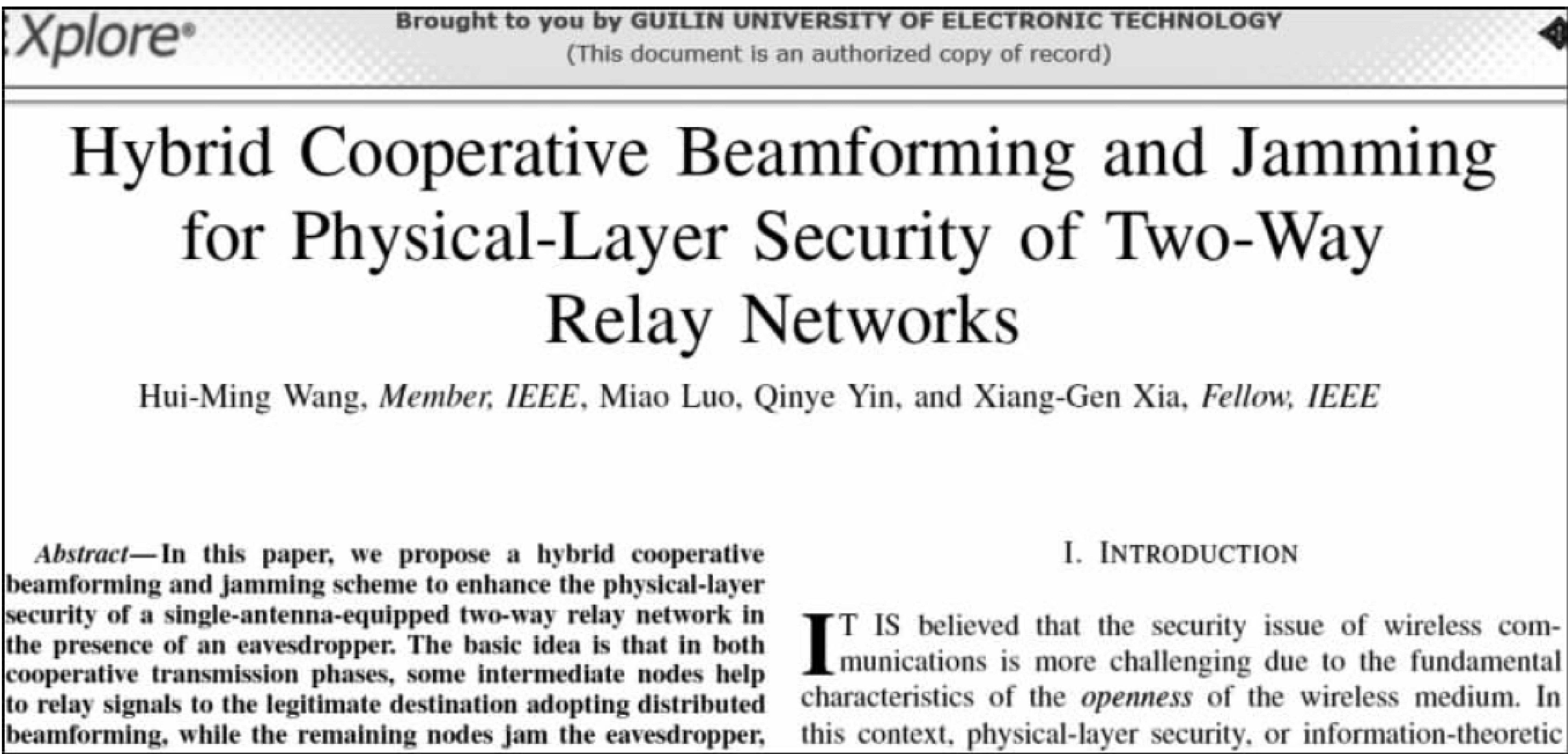


图 13-55　IEL 数据库高级检索实例二

从图 13-55 中可以看出排在第一的学术论文被引用次数为 41 次,然后进一步查看其英文,可以用网页或 PDF 两种格式查看原始全文内容,查看原文的实例如图 13-56 所示。


Xplore
Brought to you by GUILIN UNIVERSITY OF ELECTRONIC TECHNOLOGY
(This document is an authorized copy of record)

Hybrid Cooperative Beamforming and Jamming for Physical-Layer Security of Two-Way Relay Networks

Hui-Ming Wang, *Member, IEEE*, Miao Luo, Qinye Yin, and Xiang-Gen Xia, *Fellow, IEEE*

***Abstract*—In this paper, we propose a hybrid cooperative beamforming and jamming scheme to enhance the physical-layer security of a single-antenna-equipped two-way relay network in the presence of an eavesdropper. The basic idea is that in both cooperative transmission phases, some intermediate nodes help to relay signals to the legitimate destination adopting distributed beamforming, while the remaining nodes jam the eavesdropper,**

I. Introduction

IT IS believed that the security issue of wireless communications is more challenging due to the fundamental characteristics of the *openness* of the wireless medium. In this context, physical-layer security, or information-theoretic

图 13-56　IEL 数据库高级检索实例三

13.5.3 EBSCO 学术资源平台检索

(1) EBSCO 学术资源检索平台概述。EBSCO 学术资源检索平台是美国 EBSCO 公司的全文数据库检索系统，目前有全文数据库 19 种，其中最主要的有以下四种。

① 学术期刊全文数据库(Academic Source Complete，ASC)：数据库收录期刊 12 800 多种，包括 8700 多种全文期刊（其中 7613 种为专家评审期刊)，553 种非期刊类全文出版物，收录年限：1887 年至今。

② 商业资源数据库(Business Source Complete)。该数据库收录 3319 种期刊索引及摘要，其中 2300 种为全文期刊(包括 1100 多种同行评审全文期刊)及 10 000 多种非刊全文出版物(如案例分析、专著、国家及产业报告等)，收录年限：1886 年至今。

③ Communication & Mass Media Complete(CMMC，大众传媒全文数据库)。它收录著名学协会及出版社的 820 多种期刊，其中 500 种为全文收录。

④ EBSCO 电子图书(原名：NetLibrary 电子图书)。它提供 30 多万种电子图书，涉及各个主题并涵盖多学科领域。除英文电子书外，还收录法文、德文、日文和西班牙文。除提供全文的电子书外，还提供 16 000 多种有声电子图书。EBSCO eBooks 电子书可以直接进行检索，不需要安装任何阅读软件即可阅读、保存和打印，每次均可保存和打印。

EBSCO 学术资源检索平台的数据库选择见图 13-57。

(2) EBSCO 学术资源基本检索。直接输入检索词即可，下面是用“Network intrusion detection”(网络入侵检测)为检索词的基本检索结果，包括发挥结果论文总数、相关性排序等内容。见图 13-58。

(3) EBSCO 学术资源高级检索。为了便于精确检索，用户可以根据需要增加或减少检索词的输入数量，而且可以对每一个检索项设定检索字段(所有文本、作者、标题等)，同时选择布尔逻辑、检索的位置、返回结果的日期等丰富的高级检索功能。见图 13-59。

13.5.4 Wiley 在线图书馆检索

Wiley 出版商于 1807 年创立于美国，是全球历史最悠久、最知名的专业学术出版商之一，享有世界第一大独立的学术图书出版商和第三大学术期刊出版商的美誉。Wiley 在线图书馆建设了世界上最广泛的多学科在线服务数据库，包括农业、工业、建筑、化学、商业与经济、生命与健康、计算机科学、物理科学、环境科学、宇航、数学与统计学、心理学等社会与人文科学，提供访问的资源总量超过 600 万篇文章(资源来自 1500 多种期刊)，以及 18 000 本在线图书、数百本参考书、实验室指南和数据库。Wiley 在线图书馆由于其丰



图 13-57　EBSCO 学术资源检索平台的数据库选择



图 13-58　EBSCO 学术资源检索平台的基本检索实例



图 13-59 EBSCO 学术资源检索平台的高级检索视图

富的学科资源属性，我国很多高校图书馆都购买了全部或部分在线资源。

（1）Wiley 在线图书馆一般检索。一般检索默认在全部资源（all content）中检索，也可以选择具体的出版刊物名称（publication titles）即在具体的资源库中检索，具体的出版物资源名称可以在页面顶部模块 Publications 中查看，也可以查看其目录数据库（browse by subject）的具体内容。见图 13-60。

（2）Wiley 在线图书馆高级检索。在高级检索中，默认为三个检索词，可以根据需要用 Add another row 来增加检索项，检索词之间依然是典型的布尔逻辑关系（与、或、非）组配，同时可以限制信息的时间范围和检索词的位置限定（文章标题、全文、全部字段等）。图 13-61 是用三个检索词的检索实例。

图 13-62 是用三个检索词 page、ranking、algorithm 进行逻辑组配“page in Article Title AND ranking in FullText NOT algorithm in All Fields”所获得的检索结果。

13.5.5 其他典型期刊学术论文检索系统

1. SpringerLink 电子期刊

德国施普林格（Springer-Verlag）是世界上著名的科技出版集团，通过 SpringerLink

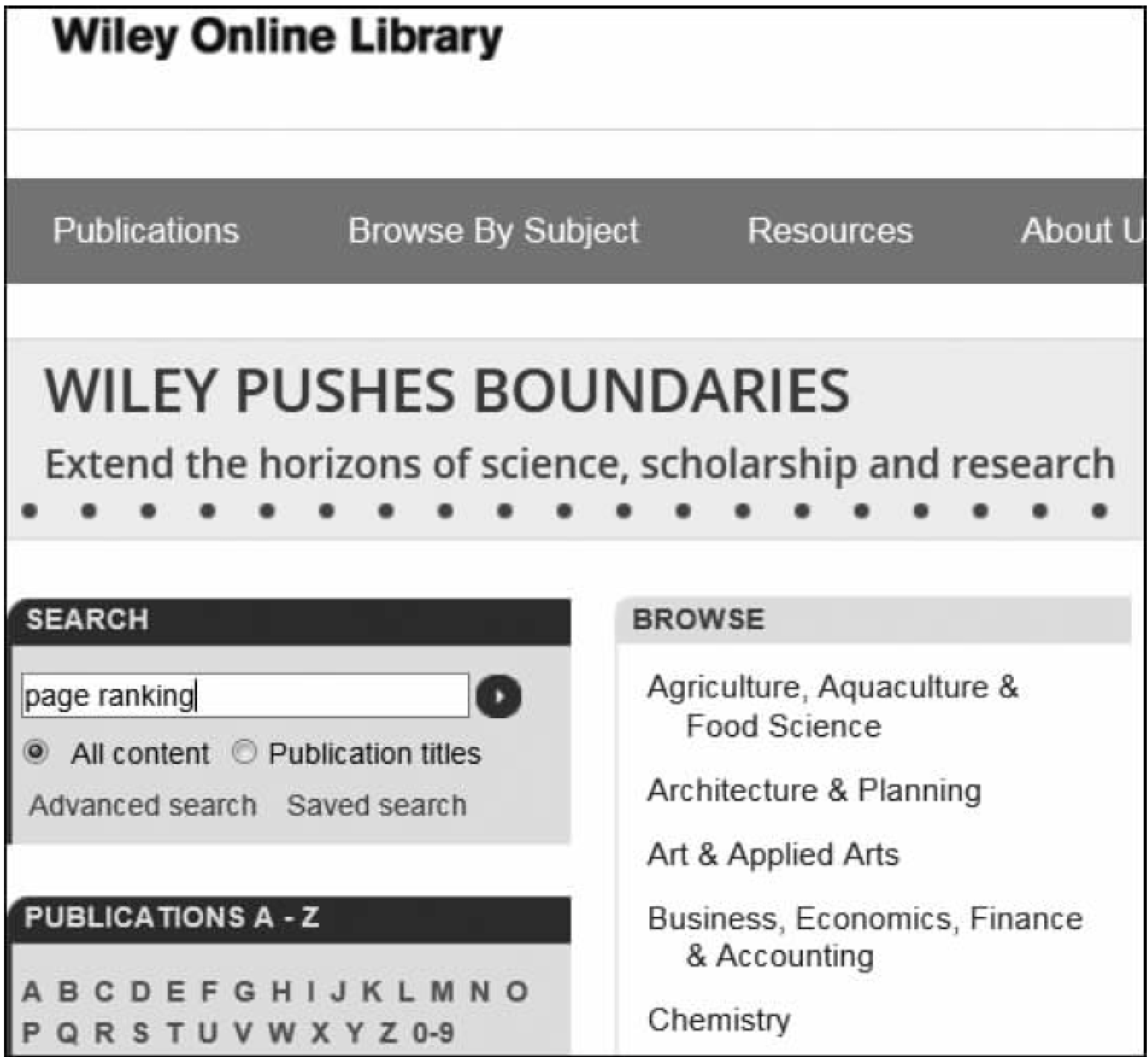


图 13-60　Wiley 在线图书馆一般检索视图

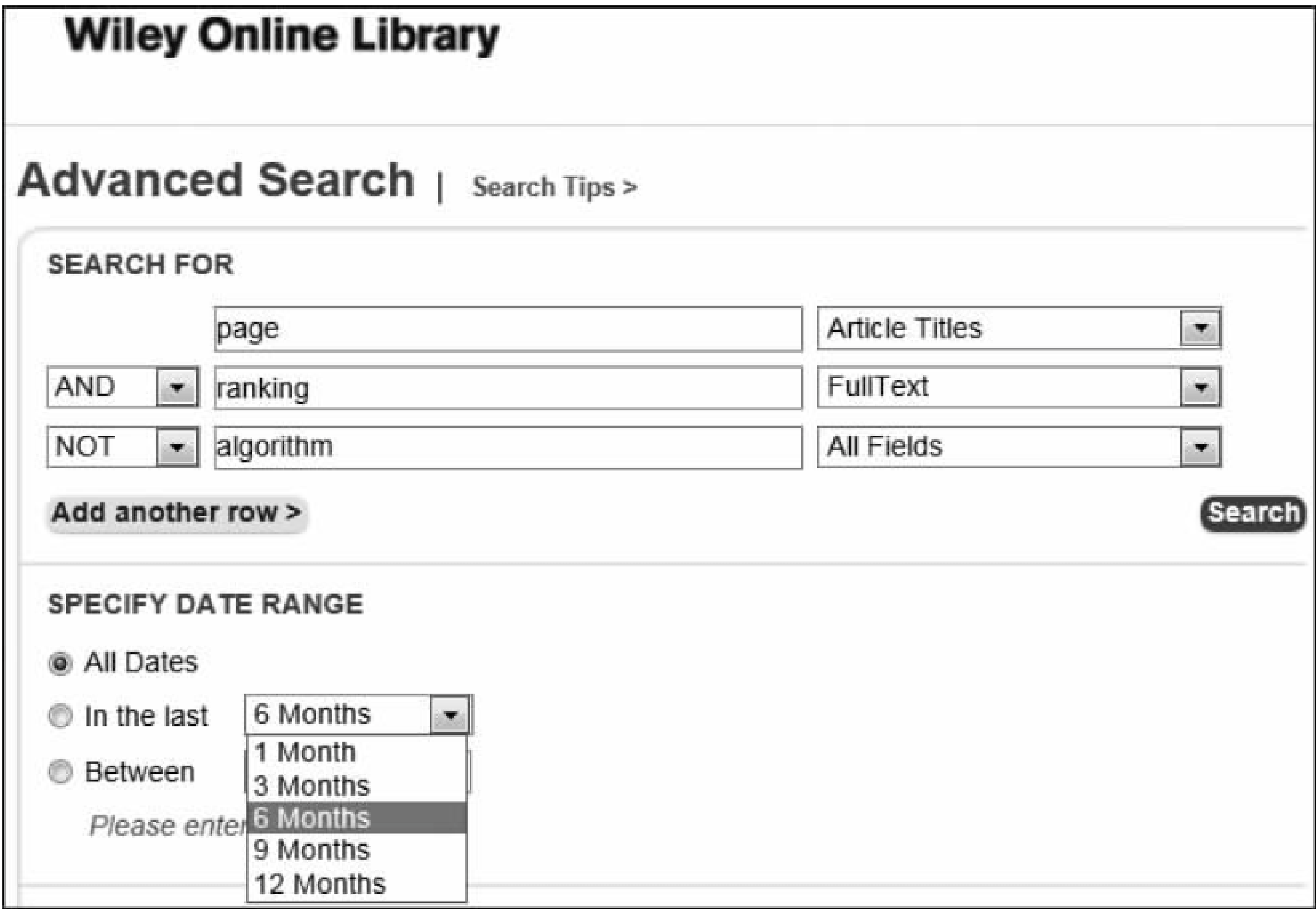


图 13-61　Wiley 在线图书馆高级检索实例一

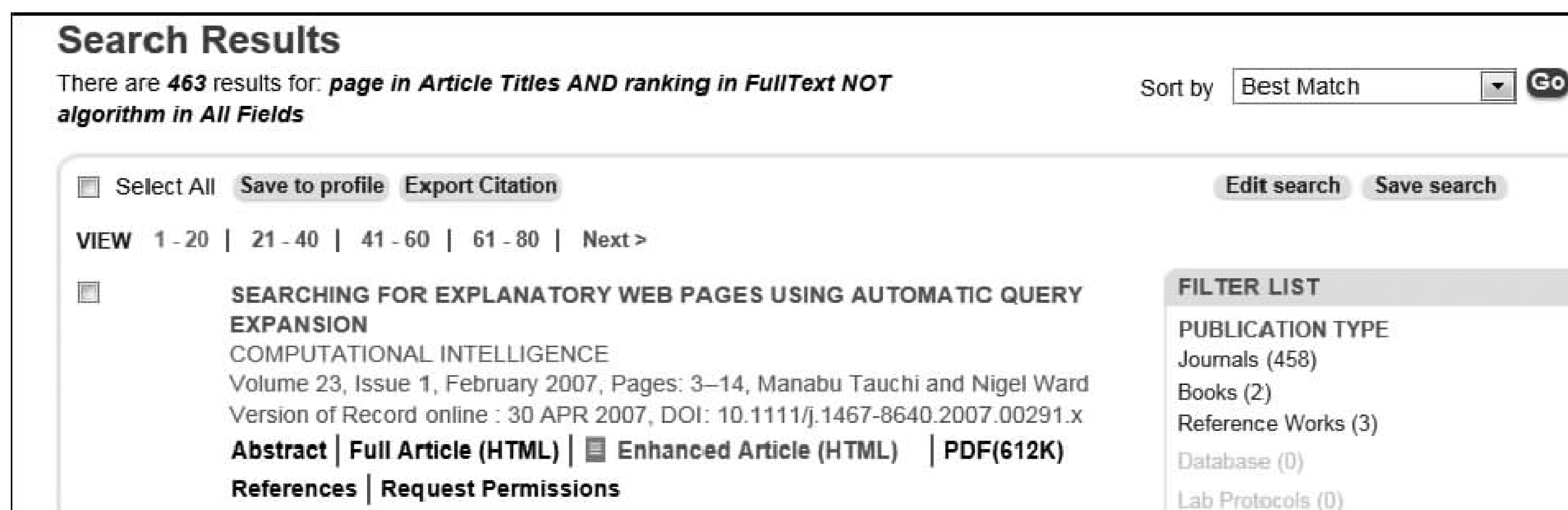


图 13-62　Wiley 在线图书馆高级检索实例二

系统提供其学术期刊及电子图书的在线服务，这些期刊是科研人员的重要信息源。2002 年 7 月开始，Springer 公司在国内开通了 SpringerLink 服务。SpringerLink 所有资源划分为 12 个学科：建筑学、设计和艺术；行为科学；生物医学和生命科学；商业和经济；化学和材料科学；计算机科学；地球和环境科学；工程学；人文、社科和法律；数学和统计学；医学；物理和天文学。原 Kluwer 出版集团出版的电子期刊已合并至该平台，另外也可以通过 Kluwer 本地服务器进行访问。SpringerLink 电子期刊目前的期刊论文资源总量达到 580 多万篇。

2. ProQuest 学术期刊数据库

学术研究图书馆(Academic Research Library，ARL)：综合参考及人文社会科学期刊论文数据库，收录近 4000 种综合性期刊和报纸的文摘/索引，其中 2365 种是全文期刊，可检索 1971 年以来的文摘和 1986 年以来的全文。

数据库涵盖的学科包括商业与经济、教育、保护服务/公共管理、社会科学与历史、计算机、科学、工程/工程技术、传播学、法律、军事、文化、医学、卫生健康及其相关科学、生物科学/生命科学、艺术、视觉与表演艺术、心理学、宗教与神学、哲学、社会学及妇女研究等领域。

3. SAGE 全文电子期刊

SAGE 公司于 1965 年成立于美国，最初以出版社会科学类学术出版物起家，自 1995 年以来，也开始陆续出版科学、技术、医学(STM)三大领域的文献。至今为止已经与 180 多家专业的学术协会和组织建立了紧密的合作伙伴关系(主要为欧美协会和组织)。目前 SAGE 连续出版高品质学术期刊 460 多种，每年出版 12～15 种百科全书和 500 余种新书。SAGE 出版的学术期刊为 100%同行评审，其中 46%的期刊被 2005 年的 Thomson Scientific Journal Citation Report(SSCI 以及 SCI)收录，另有 51 种在其所在学科类别中

排名在前十位。

SAGE Premier：包含 SAGE 出版的 452 种高品质学术期刊全文，涉及社会及人文科学、医药、科技理工等 40 个学科。收录年限是 1999 年至今，访问平台为 SAGE Journals Online(SJO)。用户访问地址为：http://online.sagepub.com/。

SAGE Deep Backfile：包含 SAGE 出版的 300 多种高品质学术期刊的全文(过刊)，收录年限是该期刊的第 1 卷第 1 期(如有)至 1998 年，访问平台为 SAGE Journals Online (SJO)。用户访问地址为：http://online.sagepub.com/。

4. Russian Library of Science——俄罗斯在线科学图书馆

俄罗斯在线科学图书馆(RLoS)囊括了俄罗斯及前独联体国家最高水平的学术机构和学协会近年来所发表的最高水平的文章和期刊论文，所译期刊全部经同行评议、专家翻译。通过 SpringerLink 平台提供 200 多种英文版俄罗斯科学期刊，其中的 100 多种来自 MAIK Nauka 出版社(是著名的俄罗斯科学院御用出版社)，另有 45 种重要科技期刊来源于 Allerton Press 出版社(自 2005 年 1 月起收录)，有 144 种期刊为 JCR 来源刊。

5. Psychology & Behavioral Sciences Collection

Psychology & Behavioral Sciences Collection 是一个综合型数据库，包含有关精神和行为特征、精神病学和心理学、心理过程、人类学以及观察和实践方法的信息。它是世界上最大的全文心理数据库，收录了 563 种期刊的全文。

6. WorldSciNet 电子期刊

WorldSciNet 为新加坡 World Scientific Publishing Co. 电子期刊发行网站，目前提供 107 种全文电子期刊，涵盖数学、物理、化学、生物、医学、材料、环境、计算机、工程、经济、社会科学等领域。

7. LWW 医学电子期刊全文数据库

OVID Technologies 公司是世界著名的数据库提供商，于 2001 年 6 月与美国银盘(SilverPlatter Information)公司合并，组成全球最大的电子数据库出版公司。目前包含生物医学的数据库有临床各科专著及教科书、循证医学、MEDLINE、EMBASE 以及医学期刊全文数据库等。OVID 全文期刊库(Journals@Ovid)提供 60 多个出版商出版的科学、技术及医学期刊 1000 多种，其中包括 Lippincott，Williams & Wilkins 出版社出版的期刊。

Lippincott，Williams & Wilkins(LWW)是世界上第二大医学出版社，其临床医学及护理学尤为突出。LWW 电子期刊全文数据库收录 235 种医学期刊，其中 154 种为核心刊(90% 为英、美核心刊)，约 150 种刊被 ISI 收录，且影响因子较高。回溯期最早至

1993 年。

8. Kluwer Online 电子期刊

荷兰 Kluwer Acdemic Publisher 是具有国际性声誉的学术出版商，它出版的图书、期刊一向品质较高，备受专家和学者的信赖和赞誉。Kluwer Online 是 Kluwer 出版的 800 余种期刊的网络版，专门基于互联网提供 Kluwer 电子期刊的查询、阅览服务。

面向 CALIS 院校提供服务的 Kluwer Online 镜像服务站在北京大学图书馆建立并开通，通过该镜像站，高校师生用户可以继续使用 Kluwer Acdemic Publisher 的 800 种电子刊，免费进行检索、阅览和下载全文。Kluwer Online 电子期刊涵盖 20 多个学科专题：Biological Sciences（73 种）、Law（59 种）、Medicine（71 种）、Psychology（57 种）、Physics（14 种）、Philosophy（35 种）、Astronomy（7 种）、Education（22 种）、Earth Sciences（18 种）、Linguistics（8 种）、Mathematics（33 种）、Social Sciences（37 种）、Computer Sciences（35 种）、Business Administration（15 种）、Engineering（19 种）、Management Science（4 种）、Electrical Engineering（13 种）、Archaeology（5 种）、Materials Sciences（13 种）、Humanities（2 种）、Environmental Sciences（8 种）、Chemistry（23 种）。

9. HeinOnline 法律全文数据库

HeinOnline 法律数据库是美国著名的法律全文数据库（网址：www.heinonline.org），涵盖全球最具权威性的近 1300 种法律研究期刊，同时还包含 675 卷国际法领域权威巨著，100 000 多个案例，1000 多部精品法学学术专著和美国联邦政府报告全文等。该数据库所收录的期刊是从创刊开始，大多数资源已更新到前一年，是许多学术期刊回溯查询的重要资源，曾获得国际法律图书馆协会（IALL）、美国法律图书馆协会（AALL）等颁发的奖项。

10. Cambridge Journals Online

剑桥大学出版社（Cambridge University Press，CUP）成立于 1514 年，是世界上历史最悠久的出版社。该社出版 220 多种学术期刊，涉及自然科学、人文社会科学及医学各个学科，大部分期刊网络版回溯到 1997 年。

2008 年，剑桥大学出版社出版 223 种学术期刊，其中 134 种人文社科类期刊，105 种自然科学类期刊，有 17 种文理交叉的期刊。总计有 132 种期刊被 Web of Science 收录，SCIE 收录 61 种、SSCI 收录 45 种、A&HCI 收录 38 种，其中有 16 种刊被重复收录。以下根据国内大学的教学和科研情况，分为自然科学（STM）、人文社科（HSS）、医学（Medicine）和工程（Engineering）四大学科数据库。

（1）自然科学类：总计 105 种，其中 63%被 SCI 收录，学科包括数学、物理、农学、生

命科学、动植物学、计算机科学、地球和大气学、科学史等。其中以数学、环境与保护生物学、农业、神经学与心理学见长。

(2) 人文社科类：总计 134 种，其中 55%的期刊被 SSCI 或 A&HCI 收录。学科包括历史、地域研究 、英语语言学等。其中以地域研究、历史、政治学和语言学见长。

(3) 医学类：剑桥大学出版社总计有 47 种医学期刊，其中 28 种被 SCI 收录，占总数的 60%。其中神经学和营养学非常出色。

(4) 工程技术类：总计有 39 种期刊，其中有 25 种被 SCI 收录，占总数的 65%。

本 章 小 结

图书是以传播知识为目的，用文字或其他信息符号记录于一定形式的材料之上的著作物；图书是人类社会实践的产物，是一种特定的不断发展着的知识传播工具。图书的基本构成要素有被传播的知识信息、有记录知识内容的文字或图像的信号、有存储与传播知识信息的物质载体、有图书的特定生成技术和工艺。图书的含义十分丰富，图书一般指书籍，由出版社出版的相对独立的出版物；有特定的书名和著(编)者名；每种书有不同的篇幅(印张)和不同的定价，并标有国际图书标准书号 ISBN。图书主要分为社会科学和自然科学两大类。本章所指的是其狭义概念即书籍，即大学生能够通过图书馆或网络查询并获取的纸质与数字化图书。

期刊也称杂志，是定期或不定期的连续出版物。每期版式基本相同，有固定名称。用卷、期或年、月顺序编号出版，有专业性和综合性两大类。期刊是由杂志社定期出版的连续出版物，如半月刊、月刊、双月刊和季刊等。刊物有固定的名称、固定的印张和固定的定价，并使用国际标准期刊号(连续出版物号)ISSN；可设有多个栏目，版式比较活泼，内容包罗万象，并可做广告。刊物出版后一般不重印，但可制作合订本。期刊内容一般比较杂，故又称杂志，期刊分专业性和综合性两大类。本书所指的期刊是对大学生的自主学习、协作学习、探究性学习有辅助作用的学术期刊。

图书与期刊的主要区别是期刊使用的是 ISSN (International Standard Serial Number，ISSN)，即国际标准期刊号，俗称连续出版物号。图书使用的是 ISBN (International Standard Book Number，ISBN)，即国际标准图书号。

本章主要以检索应用与检索实例为主，重点说明了大型中文图书目录检索系统(中国国家图书馆联机公共目录查询系统、CALIS 联合目录公共检索系统、北京大学图书馆公共查询系统和清华大学图书馆馆藏目录检索系统)和典型中文数字图书(即超星数字图

书)的检索应用。着重阐述了典型中文学术期刊论文检索系统(CNKI 中国学术期刊网和维普中文科技期刊)以及典型外文电子图书检索系统(CADAL 外文图书、世界电子图书馆、ebrary 电子图书馆和 OCLC FirstSearch 等)的检索应用。进一步阐述了典型外文学术期刊检索系统的检索应用,包括 Web of Science、IEL、EBSCO 学、Wiley、SpringerLink 和 ProQuest 等学术期刊网络数据库的检索应用。

本章思考与练习题

1. 举例说明你所在高校图书馆的主要图书资源馆藏内容与特色。
2. 举例说明你所在高校图书馆的主要学术期刊资源馆藏内容与特色。
3. 什么是图书和学术期刊? 分别有哪些主要类型?
4. 有哪些主要国内大型中文图书目录检索系统?
5. 有哪些主要国外大型外文图书目录检索系统?
6. 举一个实例说明你所在高校图书馆馆藏目录检索系统的一般检索应用方法。
7. 结合你自身的专业信息需求实际,说明中国国家图书馆联机公共目录查询系统的高级检索结果。
8. 举例说明中国国家图书馆联机公共目录查询系统的通用命令语言检索应用。
9. 举例说明 CALIS 联合目录公共检索系统的一般检索方法。
10. 举例说明 CALIS 联合目录公共检索系统的高级检索方法。
11. 举例说明清华大学图书馆图书公共查询系统与北京大学图书馆公共查询系统的异同点。
12. 举例说明典型中文数字图书检索——超星数字图书馆的多种检索功能。
13. 典型中文学术期刊论文检索平台有哪些?
14. 举一实例说明典型中文学术期刊论文检索系统的高级检索功能的应用查准率如何。
15. 典型外文电子图书检索系统有哪些?
16. 典型外文电子学术期刊检索系统有哪些?
17. 举例说明 OCLC FirstSearch 的电子图书与电子期刊检索的差异。
18. 举例说明 IEL 数据库检索与 Web of Science 在高级检索方面的异同。

参 考 文 献

[1] 吴晓兵.大学生科研创新与信息素养[M].北京：北京理工大学出版社，2013.

[2] 宋凯.大学生信息素养教程[M]. 北京：国防工业出版社 ，2013.

[3] 王吉庆.信息素养论[M].上海：上海教育出版社，1999.

[4] 唐伦刚，储冬红.大学生信息素养教育[M]. 武汉：华中科技大学出版社，2015.

[5] 何高大."美国高等教育信息素养能力标准"及其启示[J].现代教育技术，2002(3).

[6] 刘跃华. 大学生信息素养培养的影响因素及对策研究[D]. 南宁：广西大学，2014.

[7] 薛波波. 外语类院校大学生信息素养现状调查研究[D].西安：西安外国语大学，2015.

[8] 洪星.论师范大学生信息素养的培养[J].阜阳师范院学报(社会科学版)，2004(3).

[9] 赵雅萍.大学生信息素养评价指标体系构建及应用研究[D]. 济南：山东大学，2013.

[10] 王艳博.大学生信息素养的现状及培育途径研究[D].长春：东北师范大学，2009.

[11] 侯硕知.大学生信息素养课程建设研究[D].沈阳：沈阳师范大学，2011.

[12] 周芳笃.试论高校图书馆与大学生信息素养教育[J].农业图书情报学刊，2011(4).

[13] 田斌.大学生信息素养教育课程体系的构建[J].科技创新导报，2011(8).

[14] 李丽萍，韩庆年.构建教学信息管理平台提高学生信息素养[J].网络科技时代，2002(1).

[15] 李丹，刘大伟.大学生学术不端行为的学术生态思考[J].理论月刊，2015(12).

[16] 埃里希・弗洛姆.健全的社会[M].北京：中国文联出版公司，1988.

[17] 埃里希・弗洛姆.逃避自由[M].北京：工人出版社，1987：161.

[18] 马尔库塞.单向度的人[M].重庆：重庆出版社，1993：6，8.

[19] 张丹，等.防止大学生学术不端行为频发的对策研究[J].黑河学刊，2013(12).

[20] 本・阿格尔.西方马克思主义概论[M].北京：中国人民大学出版社，1991.

[21] 刘韵涵.知识产权管理[M].昆明：云南大学出版社，1997.

[22] 马海群.知识产权与信息管理[M].哈尔滨：黑龙江人民出版社，1997.

[23] 富田彻男.市场竞争中的知识产权[M].北京：商务印书馆，2000.

[24] 叶京生.知识产权与世界贸易[M].上海：立信会计出版社，2002.

[25] 沈固朝，施国良.信息源和信息采集[M].北京：清华大学出版社，2012.

[26] 范新容.大数据环境下高校图书馆提升大学生信息素养的思考与探索[J].大学图书情报学刊，2014(6).

[27] 张明海，龙献忠.云传播时代大学生信息素养教育创新研究[J].图书馆，2014(5) .

[28] 杨虎民，余武. 当代大学生信息素养的现状调查与思考[J].教育研究与实验，2014(2).

[29] 潘燕桃，廖昀赟.大学生信息素养教育的"慕课"化趋势[J]. 大学图书馆学报，2014(4).

[30] 冯婧.对大学生信息素养现状的分析及思考[J].图书馆工作与研究，2014(4).

[31] 谢群,潘宏,张俊.大学生信息素养教育的创新与思考[J].高校图书馆工作,2014(2).
[32] 王友富.普及大学生信息素养教育,提振图书馆学学科地位[J].大学图书馆学报,2014(2).
[33] 翟爱玲,粟蕾悦.大学生信息素养教育与文献检索课程建设[J].农业图书情报学刊,2013(11).
[34] 文炯.广州地区大学生信息素养水平调查研究[J].高校图书馆工作,2013(2).
[35] 路强,刘颖.大学生信息素养现状分析与培养途径研究[J].情报科学,2013(10).
[36] 沈萍,等,高校大学生学术不端行为的防治研究[J].长春教育学院学报,2014(8).
[37] 傅旭波,吴明证.道德自我调节视角下的学业自我效能感与大学生学术不端的关系研究[J].应用心理学,2013(4).
[38] 王斌.信息检索导论[M].北京:人民邮电出版社,2010.
[39] 布切尔,等.信息检索:实现和评价搜索引擎[M].陈建,等,译.北京:机械工业出版社,2011.
[40] 曼宁拉哈万.信息检索导论[M].北京:人民邮电出版社,2010.
[41] 孙建军.信息检索技术[M].北京:科学出版社,2004.
[42] 巴伊赞-耶茨.现代信息检索[M].北京:机械工业出版社,2005.
[43] 饶安平.科技信息检索[M].成都:四川科学技术出版社,2008.
[44] 李跃珍,等.信息检索与利用[M].浙江:浙江大学出版社,2006.
[45] 郝风素,等.信息资源组织与检索[M].北京:机械工业出版社,2005.
[46] 肖明.基于内容的多媒体信息索引与检索概论[M].北京:人民邮电出版社,2009.
[47] 苏新宁.信息检索理论与技术[M].北京:科学技术文献出版社,2004.
[48] 赵子江.多媒体技术基础[M].北京:机械工业出版社,2009.
[49] 焦淑红.多媒体信息系统[M].北京:机械工业出版社,2007.
[50] 刘永.多媒体信息处理[M].北京:中国农业大学出版社,2005.
[51] 刘挺.信息检索系统导论[M].北京:机械工业出版社,2008.
[52] 吴赣昌.概率论与数理统计[M].北京:中国人民大学出版社,2009.
[53] 焦玉英,温有奎,陆伟.信息检索新论[M].武汉:武汉大学出版社,2008.
[54] 赖金福,王冲,等.现代科技信息检索[M].西安:西安电子科技大学出版社,2000.
[55] 袁津生,赵传刚.搜索引擎与信息检索教程[M].北京:中国水利水电出版社,2008.
[56] 符绍宏.信息检索[M].北京:高等教育出版社,2004.
[57] 吴新博.现代信息检索简明教程[M].北京:清华大学出版社,2006.
[58] 郭太敏.信息资源检索与利用[M].北京:中国矿业大学出版社,2003.
[59] 刘廷元,邵卫东,汤凝.信息检索教程[M].北京:北京交通大学出版社,2008.
[60] 陆建江,张亚非,等.智能检索技术[M].北京:科学出版社,2009.
[61] 王玉.信息资源检索与利用[M].北京:中国人民大学出版社,2011.
[62] 郭仕明.现代信息检索[M].哈尔滨:黑龙江教育出版社,2004.
[63] 夏立新.信息检索原理与技术[M].北京:科学出版社,2009.

[64] 苏新宁.信息检索理论与技术[M]. 北京：科技文献出版社，2004.
[65] 陈雅芝.信息检索[M].北京：清华大学出版社，2006.
[66] 叶继元.信息检索导论[M].第 2 版.北京：电子工业出版社，2009.
[67] 张文修，梁怡.遗传算法的数学基础[M].西安：西安交通大学出版社，2003.
[68] 张海政.信息检索[M]. 合肥：合肥市科学技术出版社，2007.
[69] 文德.信息检索[M]. 福州：福州科学技术出版社，2007.
[70] 吴延熊.信息检索教程[M].北京：中国传媒大学出版社，2010.
[71] 瞿国忠.查询扩展技术研究[D].武汉：华中师范大学，2007.
[72] 荣光.中文文本分类方法研究[D].济南：山东师范大学，2009.
[73] 李荣陆.文本分类及其相关技术研究[D].上海：复旦大学，2005.
[74] 周龙.基于朴素贝叶斯的分类方法研究[D].合肥：安徽大学，2006.
[75] 王峻.朴素贝叶斯分类模型的研究与应用[D].合肥：合肥工业大学，2006.
[76] 程军.基于统计的文本分类技术研究[D].北京：中国科学院研生究院，2003.
[77] 余俊英.文本分类中特征选择方法的研究[D].南昌：江西师范大学，2007.
[78] 赵联冠.分布式信息检索引擎的分析与实现[D].上海：华东师范大学，2010.
[79] 李双庆.Web 服务器集群技术研究[D].重庆：重庆大学，2003.
[80] 张小伟.直觉模糊有穷自动机及其语言的研究[D].西安：陕西师范大学，2008.
[81] 吴芳.基于本体的跨语言全文检索模型的研究[D].北京：北京邮电大学，2005.
[82] 傅玲玲.基于模糊理论的语言建模的研究[D].南京：南京信息工程大学，2007.
[83] 苏绥绥.基于统计语言模型的跨语言信息检索[D].大连：大连理工大学，2009.
[84] 高立琦.基于语言模型的句子检索技术研究[D].哈尔滨：哈尔滨工业大学，2007.
[85] 张俊林.基于语言模型的信息检索系统研究[D].北京：中国科学院软件研究所，2004.
[86] 巢炼.基于图理论的 Web 服务发现方法研究[D].长沙：湘潭大学，2007.
[87] 李盛韬.基于主题的 Web 信息采集技术的研究[D].北京：中国科学院计算机研究所，2002.
[88] 周艳.主题 Web 信息采集系统的研究与设计[D].绵阳：西南科技大学，2008.
[89] 白田恬.基于贝叶斯网络的信息检索研究[D].重庆：重庆大学，2007.
[90] 王薇.基于内容的音频检索特征提取技术研究[D].上海：上海交通大学，2008.
[91] 李晓光，王大玲，于戈.基于统计语言模型的信息检索[D].沈阳：东北大学信息科学与工程学院，2005.
[92] 赵夷平.传统搜索引擎与语义搜索引擎比较研究[D].长春：吉林大学，2009.
[93] 吴学义.基于 web service 的企业搜索引擎的构架及优化[D].长春：吉林大学，2008.
[94] 张伟.垂直搜索引擎的设计与实现[D].西安：西安电子科技大学，2008.
[95] 张纯青.科技文献检索系统相关排序算法研究[D].合肥：中国科学技术大学，2007.
[96] 陈鹏.基于统计的搜索引擎中文输入纠错技术研究[D].北京：北京邮电大学，2010.

[97] 洪颖.面向化工领域的智能元搜索引擎系统的研究与设计[D].北京：北京化工大学，2008.
[98] 张毅波.中文结构化信息检索系统的研究与实现[D].北京：中国科学院软件研究所，2001.
[99] 邱春艳.基于粗糙集理论的智能信息检索方法的研究[D].长春：东北师范大学，2005.
[100] 权小军.基于潜在语义结构的文本层次分类[D]. 合肥：中国科学技术大学，2008.
[101] 金玉坚.基于层次搜索的信息过滤和检索方法研究[D].上海：上海大学，2005.
[102] 邓剑勋.信息检索中的相关反馈技术研究[D].重庆：重庆大学，2007.
[103] 张静.基于粗糙集理论的数据挖掘算法研究[D].西安：西北工业大学，2006.
[104] 黄治国.基于粗糙集的信息检索研究[D].长沙：中南大学，2007.
[105] 高珊.信息检索中的查询扩展及相关技术研究[D].武汉：华中师范大学，2008.
[106] 王梁.基于向量空间的信息检索算法研究[D].长春：长春理工大学，2007.
[107] 李媛媛.基于潜在语义索引的中文文本检索研究[D].成都：西南交通大学，2008.
[108] 吴桂宾.基于神经网络的网页排序学习算法研究[D].广州：中山大学，2009.
[109] 朱敏.基于自适应遗传 BP 神经网络的文本分类方法[D].南昌：南昌大学，2010.
[110] 赵中原.基于神经网络的中文文本分类研究[D].南京：南京邮电大学，2005.
[111] 张龙.基于粗糙集和神经网络的中文文本分类研究与实现[D].西安：西北大学，2008.
[112] 邱兴兴.基于模糊逻辑和神经网络的文本分类方法[D].南昌：南昌大学，2007.
[113] 吴琼.基于神经网络的词的切分及切分歧义消解[D].长春：长春工业大学，2007.
[114] 杨哲.提高信息检索性能的有效机制与算法研究[D].北京：中国科学院研究生院，2004.
[115] 陈定权.信息检索系统中的用户相关反馈机制[J].现代图书情报技术，2002(4).
[116] 杨俊柯，杨贯中，杨建学.基于语义模型的信息检索机制研究[J].计算机工程，2006(12).
[117] 丁国栋，白硕，王斌.文本检索的统计语言建模方法综述[J].计算机研究与发展，2006(5).
[118] 金迪，马衍民.PageRank 算法的分析及实现[J].经济技术协作信息，2009(18).
[119] 闫泼，马军.面向主题的网页排序算法研究[J].第三届全国信息检索与内容安全学术会议，2007(3).
[120] 姜鑫维，赵岳松.Tapic PageRank——一种基于主题的搜索引擎[J].计算机技术与发展，2007(17).
[121] 王丽晖.WEB 页面信息采集技术[J].网络技术，2007(4).
[122] 郑亚，谢琳.DNS 的原理及其应用[J]. 软件导刊，2012(6).
[123] 谢大吉.基于 Java 的网络制造资源主题信息采集模块设计[J]. 计算机工程与设计，2010 (19).
[124] 曾义聪，杨贯中，周志光，等. 基于层次语义的 URL 排序方法研究[J]. 计算机工程与设计，2008(13).
[125] 杜欢.主题 Web 信息采集技术[J].四川理工学院学报（自然科学版），2007(5).
[126] 陈飞等.基于 HITS 算法的查询结果多样化方法[J].山东大学学报(理学版)，2011(5).
[127] 郭少友. Web 环境下分布式信息检索模式[J].情报科学，2003(6).

[128] 于海波.分布式索引的研究与应用[J].科技信息,2010(26).
[129] 李盛韬,余智华. web 信息采集研究进展[J].计算机科学,2003(2).
[130] 王伟.基于 Hadoop 的分布式索引集群的研究[J].电脑知识与技术,2011(7).
[131]丁国栋,王斌.Web 超链挖掘:中国境内 Web 图结构研究[J].计算机工程,2005(7).
[132] 张路, 袁晓洁.大规模数据集的分布式索引机制研究[J]. 微电子学与计算机,2008(10).
[133] 李广丽.基于网页内容评价和 Web 图的启发式垂直搜索策略的设计[J].信息系统,2009(9).
[134] 张敏,李锋.PageRank 算法研究[J].网络与通信,2011(8).
[135] 李文东.基于 WEB 的智能信息采集及处理系统研究[J].科技创新导报,2008.
[136] 张健,宋刚.基于分裂式 K 均值聚类的图像分割方法[J]. 计算机应用,2011(2).
[137] 陈韬伟.基于灰关联测度的分裂式层次聚类算法[J]. 西南交通大学学报,2010(2).
[138] 罗宏.基于查询相关性分析的检索结果聚类算法[J].小型微型计算机系统,2011(10).
[139] 李寒.基于凝聚式信息瓶颈的加权层次聚类算法[J].计算机工程,2011(6).
[140] 黄永锋,刘同明.聚集式聚类分析方法及其应用[J].华东船舶工业学院报,2002(2).
[141] 靳延安,刘行军.一种改进的层次聚类算法[J]. 武汉理工大学学报,2011(6).
[142] 程丽,周亚建.一种改进的支持向量聚类算法[J].北京电子科技学院学报,2011(6).
[143] 谢振平,王士同.一种基于软边界球分的分裂式层次聚类算法[J].模拟识别与人工智能,2008(4).
[144] 陈利军.常用的聚类技术分析[J].湖南工业职业技术学院学报,2012(12).
[145] 向培素.聚类算法综述[J].西南民族大学学报,2011(37).
[146] 刘洋.聚类算法的研究[J].长春师范学院学报,2012(31).
[147] 李洁,陈围,谭立地.基于聚类技术的学科信息检索服务[J].江西科学,2012(30).
[148] 赵妍,赵学民.基于 CURE 的用户聚类算法研究[J].计算机工程与应用,2012(48).
[149] 杨柳,张俊芝.浅谈聚类算法及其存在的问题[J].产业与科技论坛,2012(11).
[150] 王千,王成,冯振元,等.K-means 聚类算法研究综述[J].电子设计工程,2012(20).
[151] 陈晓春.基于 K-Means 和 EM 算法的聚类分析[J].福建电脑,2009(2).
[152] 陈德军,罗金成,张兵.基于改进的 K-means 聚类算法的分类评价方法[J].武汉理工大学学报,2011(3).
[153] 陶树平.多媒体数据库的数据模型研究[J].上海铁道大学学报,2003 (8).
[154] 邓佩珍.数字图书馆关键技术——数据压缩的原理与方法[J]. 图书馆学研究,2008(19).
[155] 谌群芳.多媒体内容检索技术[J].情报检索,2003(6).
[156] 王娣.多媒体数据库技术综述[J].情报杂志,2001(11).
[157] 张玫.多媒体检索:从基于内容的方法到基于概念的方法[J]. 情报杂志,2006(6).
[158] 孟倩,王松.多媒体数据库管理系统实现途径的研究[J].徐州师范大学学报(自然科学版),2000 (1).
[159] 罗菁,等.基于内容的多媒体检索和索引的研究[J].中原工学院学报,2004(1).

[160] 沈燕,等.基于内容的多媒体检索技术在数字档案馆中的应用[J].情报杂志,2004(4).
[161] 徐建华.一种新型的多媒体检索技术[J].情报学报,2000(8).
[162] 卢爱芹.基于内容的多媒体检索技术综述[J].科技传播,2010(5).
[163] 于海龙.多媒体应用中的关键技术[J].职业,2007 (21).
[164] 刘奕群,等.基于用户行为分析的搜索引擎自动性能评价[J].软件学报,2008 (11).
[165] 赵美玉.浅析计算机多媒体技术应用与发展趋势[J].信息技术,2012(5).
[166] 陈黄海,李琥,徐盛.多媒体与多媒体数据建模[J].计算机应用与软件,2002 (1).
[167] 肖健宇,张大方.多媒体数据库的关键技术:多媒体数据模型[J].计算机工程与应用,2002 (7).
[168] 陈晓燕,王克难,黄少波.浅谈多媒体数据技术的发展与应用[J].信息传媒,2010(22).
[169] 段桂英.面向对象的多媒体数据库[J].科技信息,2011(28).
[170] 岳根霞.多媒体数据库的数据模型研究[J].电脑编程技巧与维护,2012(14).
[171] 刘玉照,黄蕾.多媒体数据模型及其实现途径之比较研究[J].情报科学,2001(8).
[172] 张景春,管士学,马媛.词性分类优先在搜索引擎中的应用[J].计算机光盘软件与应用,2011(4).
[173] 冯成.个性化搜索引擎关键技术及应用[J].计算机光盘软件与应用,2011(12).
[174] 盛宪锋,山岚.基于元搜索引擎的专业式智能网络信息检索系统[J].计算机工程与设计,2004(1).
[175] 王改香.搜索引擎的体系结构与索引技术探析[J].长江大学学报,2011(3).
[176] 金波.刍议中文搜索引擎的应用技巧[J].浙江纺织服装职业技术学院学报,2005(3).
[177] 安金龙,王正欧,马振平.一种新的支持向量机多类分类方法[J].信息与控制,2004(3).
[178] 高梦娇,吕玉琴,侯宾.基于 R-tree 和倒排文件的混合索引的设计与实现[J].中国科技论文在线,2011(7).
[179] 刘凤灵.漏洞垂直搜索引擎的设计与实现[J].中国科技论文在线,2011(1).
[180] 杜亚军.智能信息处理及其在搜索引擎中的应用[J].西华大学学报,2007(3).
[181] 付雪峰,王明文.基于模糊-粗糙集的文本分类方法[J].华南理工大学学报(自然科学版),2004(32).
[182] 曾雪强.一种基于潜在语义结构的文本分类模型[J].华南理工大学学报(自然科学版),2004(32).
[183] 张浩然,汪晓东.回归最小二乘支持向量机的增量和在线式学习算法[J].计算机学报,2006(3).
[184] 李红莲,王春花,袁保宗,等.针对大规模训练集的支持向量机的学习策略[J].计算机学报,2004(5).
[185] 刘宏.一种新的支持向量机主动学习策略及其在文本分类中的应用[J].计算机科学,2003(6).
[186] 张浩然,韩正之,李昌刚.支持向量机[J].计算机科学,2002(12).
[187] 鲁松,李晓黎,白硕,等.文档中词语权重计算方法的改进[J].中文信息学报,2000(6).
[188]冯飞燕.搜索引擎:穿透互联网的动力——搜索引擎能做什么[J].电子电脑,1996(2).
[189] 张晓刚,李明树.智能搜索引擎技术的研究与发展[J].计算机工程与应用,2001(12).
[190] 邱均平,余以胜.基于知识库系统的智能搜索引擎研究[J].情报科学,2006(24).

[191] 曲卫华,王群.搜索引擎原理介绍与分析[J].开发研究与设计技术,2006(3).
[192] 杨丽萍,马继涛,张虹霞.网络搜索引擎分类与发展[J].情报学,2006(12).
[193] 李宝敏.基于本体的智能搜索引擎研究[J].情报杂志,2006(10).
[194] 颜素莉.主流中俄文搜索引擎核心技术分析与比较研究[J].计算机时代,2012(1).
[195] 刘畅.综合搜索引擎与垂直搜索引擎比较研究[J].情报学报,2007(1).
[196] 王春红,张敏隐.隐含语义索引模型的分析与研究[J].计算机应用,2007(27).
[197] 刘冰.知识库系统原理探讨[J].软件导刊,2009(8).
[198] 吕进来.智能技术在信息检索中的应用[J].计算机时代,2005(10).
[199] 黄萱箐,吴立德.基于向量空间模型的文档分类系统[J].模式识别与人工智能,1998(6).
[200] 倪廓阔,等.搜索引擎中“N1+N2”型短语查询优化研究[J].计算机应用与软件,2012(9).
[201] 袁晓丰,等.基于短语检索和答案排序的列表问题回答方法[J].中文信息学报,2008(5).
[202] 张惠文.网络信息检索技术的智能化趋势[J].情报理论与实践,2001(6).
[203] 贾红英.搜索引擎检索功能的比较研究[J].时代情报,2003(11).
[204] 陈文,李玉莲.布尔检索在 Ei-compendex 和 CNKI 中文期刊数据库中的应用区别[J].现代情报,2005(8).
[205] 谢丽聪,俞建家,张莹.布尔查询的改写算法[J].福州大学学报(自然科学版),2004(4).
[206] 俞平,肖南峰,甘志刚.第三代搜索引擎研究[J].南京信息工程大学学报:自然科学版,2009(2).
[207] 张振亚,等.基于余弦相似度的文本空间索引方法研究[J].计算机科学,2005(9).
[208] 郑继明,李瑞仙,浦兴成,等.基于单状态 HMM 的音频分类方法研究[J].计算机应用,2009(2).
[209] 刘维华,崔涛.基于内容的音频检索算法研究[J].计算机工程与设计,2006(16).
[210]郑贵滨.基于听觉模型的模糊直方图音频索引和检索方法[J].全国网络与信息安全技术讨论会,2004(8).
[211] 李晓丽,等.相容粗糙集在音频检索中的应用[J].兰州理工大学学报,2006(4).
[212] 曾柏森.基于敏感位置哈希索引的音频检索[J].程度信息工程学院学报,2009(6).
[213] 卢坚,陈毅松,孙正兴,等.基于隐马尔可夫模型的音频自动分类[J].软件学报,2002(8).
[214] 颜永红.音频信息识别与检索技术[J].China Academic Journal Electronic Publishing House,2009(3).
[215] 宋博,须德.音频信息检索的研究及实现[J].计算机应用,2003(12).
[216] 李亮,刘万春,徐泉清,等.一种基于支持向量机的专业中文网页分类器[J].计算机应用,2004(4).
[217]马金娜,田大刚.基于 SVM 的中文文本自动分类研究[J].计算机与现代化,2006(8).
[218] 徐启华,杨瑞.一种新的软间隔支持向量机分类算法[J].计算机工程与设计,2005(9).
[219] 赵丽,李天舒,刘玉蕾.基于支持向量机的机器学习的研究[J].哈尔滨师范大学自然科学学报,2008(6).
[220] 胡银霞,赵伯兴.国内外信息检索行为研究比较[J].图书馆学研究,2011(4).

[221] 龚文涛,武立莹,刘会霞,等.信息检索技术的发展概况及趋势[J].医学情报工作,2001(3).
[222] 龚蛟腾.网络信息检索技术现状、瓶颈及趋势分析[J].情报杂志,2004(5).
[223] 陈桃,明均仁.现代信息检索技术发展趋势初探[J].农业图书情报学刊,2007,19(6).
[224] 韩娇红.我国智能化信息检索发展及研究现状[J].图书馆学刊,2012(1).
[225] 孔为民.信息检索技术的新趋势[J].农业图书情报学刊,2009(3).
[226]苗兰芳，彭群生. 基于点索引的网格模型的层次结构[J].计算机辅助设计与图形学学报,2005(9).
[227] 寿涌毅,等. 基于临近性的企业网络知识转移仿真研究[J].科学学与科学技术管理,2012(1).
[228] 刘雄恩. 计算机考试软件中实现 Word 文档评分的 3 种方法[J].福建农林大学学报,2003(9).
[229] 乔亚男,齐勇,侯迪. 具有孤立项过滤的信息检索查询词的分析方法[J].西安交通大学学报,2009(8).
[230] 田萱,孟祥光,刘希玉.智能信息检索中的个性化模式的表示形式研究[J].情报学报,2004(1).
[231] 何靖. 一种问答式检索系统布尔查询生成方法[J].山东大学学报,2006(3).
[232] 杨小平,丁浩,黄都培. 基于向量空间模型的中文信息检索技术研究[J].计算机工程与应用,2003(3).
[233] 杨劲松,凌培亮.搜索引擎 PageRank 算法的改进[J].计算机工程,2009(12).
[234] 赫金隆,王成良.原创优先的搜索引擎排序算法[J].计算机工程,2008(9).
[235] 黄素珍,陈宁江,苏德富.并发多元搜索引擎的研究与应用[J].广西大学学报,2005(6).
[236]郑凯明.垂直搜索引擎应用研究[J].赤峰学院学报,2011(2).
[237] 高磊.企业搜索引擎技术研究及应用[J].计算机光盘软件与应用,2011(20).
[238] 李绍华,高文宇.搜索引擎页面排序算法研究综述[J].计算机应用研究,2007(6).
[239] 黄倩.浅谈信息检索前沿发展的几个问题[J] .大众文艺(理论),2009(19).
[240] 李红梅,等.元搜索引擎结果合成算法[J].北京邮电大学学报,2008(5).
[241] 胡伶霞,明均仁.现代信息检索技术发展探讨[J].农业图书情报刊,2009(4).
[242] 张利平.基于综合特征的图像检索技术的研究[J]. 图书馆学研究,2007(12).
[243] 刘忠伟,章毓晋.综合利用颜色和纹理特征的图像检索[J].通信学报,1999(5) .
[244] 王小玲,谢康林.一种新的基于区域的图像检索方法[J]. 计算机工程与应用,2005(3).
[245] 李向阳,庄越挺,潘云鹤. 基于内容的图像检索技术与系统[J]. 计算机研究与发展,2001(3).
[246] 崔江涛,等.基于相关反馈的高维图像检索方法[J].西安:电子科技大学学报(自然科学版),2006(2).
[247] 何立民,万跃华.数字图书馆中基于内容的图像检索关键技术[J].中国图书馆学报(双月刊),2002(6).
[248]吴锐航,李绍滋,邹丰美. 基于 SIFT 特征的图像检索[J]. 计算机应用研究,2008(2).
[249] 王涛,胡事民,孙家广. 基于颜色-空间特征的图像检索[J].软件学报, 2002(10).

[250] 许兰,李金岳. 图像元数据在检索中的应用[J]. 数字图书馆论坛,2007(12).
[251] 张玉峰,蔡昌许.基于语义的图像检索系统研究[J].中国图书馆学报(双月刊),2004(5).
[252] 陆伟,张宓,刘丹.基于 XML 文本片段的图像检索实现与评价[J].中国图书馆学报,2009(3).
[253] 霍亮,杨柳,张俊芝.贝叶斯与 k-近邻相结合的文本分类方法[J].河北大学学报(自然科学版),2012(3).
[254] 饶丽丽,等.基于特征相关的改进加权朴素贝叶斯分类算法[J]. 厦门大学学报(自然科学版),2012(4).
[255] 陈琳,王箭. 三种中文文本自动分类算法的比较和研究[J]. 计算机与现代化,2012(2).
[256] 刘冬雪.文本分类技术在信息检索中的应用[J]. 科技资讯,2010(18).
[257]高胜利.改进的朴素贝叶斯聚类 Web 文本分类挖掘技术[J].廊坊师范学院学报(自然科学版),2012(12).
[258] 袁占亭.数据抽取及语义分析在 Web 数据挖掘中的应用[J].计算机工程与设计,2005(6).
[259] 王娟,等.特征选择方法综述[J].计算机工程与科学,2005(12).
[260] 姚旭,等.特征选择方法综述[J].控制与决策,2012(12).
[261] 申红.吕宝粮,等.文本分类的特征提取方法比较与改进[J].计算机仿真 ,2006(3) .
[262] 宁慧,吕志龙.中文文本分类中特征选择方法的研究[J].电脑知识与技术,2007(11).
[263] 林海.信息检索发展浅析[J].科技情报开发与经济, 2007(10).
[264] 郭文娟.超文本检索特点研究[J].中国科技信息,2007(9).
[265] 马向东,张丽杰.加权检索与逻辑检索的比较及实现路径[J].现代图书情报技术,1995(6).
[266] 李广元,陈丹.文本信息检索技术[J].广西科学院学报,2001(5).
[267] 贺宏朝,高剑锋.一种基于上下文的中文信息检索查询扩展[J].中文信息学报,1999(5).
[268] 季春.音频信息检索技术的发展及应用[J].现代情报,2007(1).
[269] 申展,江宝林,唐磊.全文检索模型综述[J].计算机科学,2004(5).
[270] 裴飞,洪宇,孙常龙等.基于 Web 的查询扩展[J].电脑知识与技术,2011(6).
[271] 宋伟萍,杨建林.个性化信息检索中的相关反馈技术研究[J].图书情报工作,2008(4).
[272] 黄名选,严小卫,张师超.查询扩展技术进展与展望[J].计算机应用与软件,2007(11).
[273] 王志军,于超.基于隐式反馈的个人信息检索技术及实现[J].计算机工程,2003(6).
[274] 石艳霞.信息检索中“相关性”与“相关反馈”研究概述[J]. 晋图学刊,2002(2).
[275] 成全,司辉.信息检索相关性评价及其改善策略研究[J]. 情报杂志,2008(2).
[276] 苏中,张宏江,马少平.基于贝叶斯分类器的图像检索相关反馈算法[J].软件学报,2002(10).
[277] 李勇,桑艳艳.Web 智能检索中动态相关反馈技术研究[J].国际通信会议,2003(2).
[278] 李晓黎,周长胜.基于相关反馈技术的 Web 检索改进研究与实现[J]. 航空计算技术,2004(3).
[279] 石艳艳,刘南杰.相关反馈查询及其实用评价[J].计算机应用与软件,1990(3).
[280] 艾丹祥,张玉峰.相关反馈技术在知识检索中的应用[J].情报科学,2003(10).

[281] 刘绍翰，武港山，张福炎. 基于词条权值的相关反馈算法在 Web 信息检索中的应用[J]. 情报学报，2002(6).
[282]陈晓金，王兵. 信息检索扩展技术研究[J]. 图书情报工作，2008(12).
[283] 林国俊，叶飞跃，耿冬，等. 基于语义的概念查询扩展[J]. 计算机工程与设计，2009(6).
[284] 马晖男，吴江宁，潘东华. 一种基于同义词词典的模糊查询扩展方法[J]. 大连理工大学学报，2007(3).
[285] 王知津，郑红军. 基于集合理论的信息检索模型[J]. 情报科学，2004(11).
[286] 钱晴. 一种新型的扩展布尔检索模型[J]. 现代图书情报技术，1987(2).
[287] 卓佳，张俊坤，李畅. 基于向量空间模型的信息检索[J]. 华南金融电脑，2008(8).
[288] 刘斌，陈桦. 向量空间模型信息检索技术讨论[J]. 情报杂志，2006(7).
[289] 田萱，李冬梅. 上下文信息检索研究综述[J]. 计算机科学，2011(38).
[290] 朱鸽昀，李琳. 现代信息检索技术的发展概况[J]. 医学信息，1999(11).
[291] 王知津，李明珍. 十年来我国信息检索研究述评[J]. 现代图书情报技术，2004(12).
[292] 张帆，等. 迈向 21 世纪的检索技术[J]. 中国图书馆学报，1995(7).
[293] 陆承兆. 试论计算机情报检索途径和技术发展趋势[J]. 图书馆论坛，2002(3).
[294] 张颖，等. 网络信息检索展望[J]. 现代图书情报技术，2000(3).
[295] 曾民族. 文本信息检索技术进展和性能评价框架[J]. 现代图书情报技术，1997(3).
[296] 卢文林. 信息检索技术的发展概况[J]. 现代图书情报技术，2003(3).
[297]李明，王丽. 基于本体的信息检索系统模型[J]. 兰州理工大学学报，2007(2).
[298] 中国国家图书馆联机公共目录查询系统[DB/OL]. http://opac.nlc.cn.
[299] CALIS 联合目录公共检索系统[DB/OL]. http://opac.calis.edu.cn/opac/simpleSearch.do.
[300] 北京大学图书馆公共查询系统[DB/OL]. http://qjbopac.1617888.com/.
[301] 清华大学图书馆馆藏目录检索系统[DB/OL]. http://discovery.lib.tsinghua.edu.cn/.
[302] 超星数字图书馆[DB/OL]. http://www.sslibrary.com/.
[303] CNKI 中国学术期刊网检索[DB/OL]. http://cnki.hilib.com/.
[304] 维普中文科技期刊数据库[DB/OL]. http://qikan.cqvip.com/.
[305] CADAL 外文图书检索[DB/OL]. http://www.cadal.cn/.
[306] 世界电子图书馆检索[DB/OL]. http://www.ebooklibrary.org/.
[307] ebrary(电子图书馆)检索[DB/OL]. http://site.ebrary.com/.
[308] OCLC FirstSearch[DB/OL]. http://www.oclc.org/firstsearch.en.html.
[309] Early English Books Online[DB/OL]. http://eebo.chadwyck.com/home.
[310] iG Publishing[DB/OL]. http://www.igpublish.com/.
[311] Wiley Online Library[DB/OL]. http://onlinelibrary.wiley.com/.
[312] Web of Science[DB/OL]. https://apps.webofknowledge.com/.

[313] IEEE/IET Electronic Library[DB/OL]. http://ieeexplore. ieee. org/Xplore/home. jsp.
[314] EBSCO 学术资源平台[DB/OL]. http://search. ebscohost. com/.
[315] Wiley 在线图书馆[DB/OL]. http://onlinelibrary. wiley. com/.
[316] SpringerLink 电子期刊[DB/OL]. http://link. springer. com/.
[317] ProQuest 学术期刊数据库[DB/OL]. http://search. proquest. com/ebrary/index.
[318] SAGE Premier[DB/OL]. http://online. sagepub. com/.
[319] Russian Library of Science[DB/OL]. https://en. wikipedia. org/.
[320] WorldSciNet 电子期刊[DB/OL]. http://www. worldscientific. com/page/worldscinet.
[321] Kluwer Online 电子期刊[DB/OL]. http://kluwer. calis. edu. cn/.
[322] HeinOnline——法律全文数据库[DB/OL]. http://home. heinonline. org/.